


Coursework Declaration and Feedback Form

The Student should complete and sign this part

Student Number: 2448798	Student Name: Hongyu Mu
Programme of Study (e.g. MSc in Electronics and Electrical Engineering): MSc Civil Engineering	
Course Code: ENG5059P	Course Name: MSc Project
Name of <u>First</u> Supervisor: Umer Zeeshan Ijaz	Name of <u>Second</u> Supervisor: Ciara Keating
Title of Project: A Survey on Microbial Diversity in Coal Tar Contaminated Media	
Declaration of Originality and Submission Information	
<p><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></p> <p>Signed (Student) : <i>Hongyu Mu</i></p>	 E N G 5 0 5 9 P
Date of Submission : August 21, 2020	

<i>Feedback from Lecturer to Student – to be completed by Lecturer or Demonstrator</i>	
<p>Grade Awarded:</p> <p>Feedback (as appropriate to the coursework which was assessed):</p> 	
Lecturer/Demonstrator:	Date returned to the Teaching Office:

A Survey on Microbial Diversity in Coal Tar Contaminated Media

MSc Civil Engineering

Hongyu Mu (Student ID: 2448798)

Supervisor: Dr Umer Zeeshan Ijaz

Co-Supervisor: Dr Ciara Keating

August 21, 2020

**A report submitted in partial fulfillment of the requirements
for MSc Civil Engineering Degree
at the University of Glasgow**

Acknowledgements

First of all, I want to thank my two supervisors, Dr Umer Zeeshan Ijaz and Dr Ciara Keating, and I would express my heartfelt gratitude and sincere admiration to the two advisors who had been able to guide, supervise and help us wholeheartedly during their really busy schedules; it is also because of their hard work that we could be able to complete our projects both on time and smoothly. Secondly, I want to thank our university, the University of Glasgow, for providing us this opportunity to exercise, improve and surpass ourselves; and in this slightly tough period, our university had tried the best on ensuring that we could conduct the daily studying life normally, which was also the basis for us to complete our tasks. Once again, I want to thank the two schoolmates, Bozhen Chen and Keda Li, who were in the same research group with me. Your helps and various suggestions during the process of the project had benefited me a lot and given me a deeper understanding of scientific research life. Finally, I would like to express my gratitude to my family and my motherland. It is because of your silent supports and dedication that made me feel more at ease and full of conviction in my study life in a foreign country, your cares are always the greatest power of mine to make me go forward.

Abstract

Coal tar is a kind of thick dark liquid, which is a kind of byproduct of the manufactured gas industry. From the early 19th century, the manufactured gas industry, which could be regarded as a global industry, got prosperity in Europe, North America and other parts of the world; at the late 20th century, when there were plentiful natural gas fields had been discovered, the manufactured gas industry went to decline gradually.

During the process of gas manufacture, some detrimental byproducts are created, coal tar is one of them. As a kind of dense non-aqueous-phase liquids, coal tar contains massive organic and inorganic components, including some hazardous matters such as polycyclic aromatic hydrocarbons (PAHs), which could cause cancer or other health issues. These harmful matters could also be considered as a ubiquitous contaminant at former manufactured gas plant sites.

It has been estimated that more than 3000 former manufactured gas plant sites exist in the United Kingdom alone, which makes the coal tar contamination a serious environmental problem. Additionally, since the composition of the coal tar is highly dependent on the raw materials and the approach that utilized in the gas production process, and also due to the technical restriction on related fields, the issue of coal tar contaminant could not be solved effectively for a long time.

However, according to previous research, it has been pointed out that there are a number of microorganisms that could live in naturally occurring terrestrial oil seeps and natural asphalts, which are comprised of highly recalcitrant petroleum hydrocarbons. Since there are some similarities between these conditions and the soils contaminant by coal tar, it could pose a possibility of utilizing the microorganisms that live in coal tar to achieve a biodegradation of the detrimental byproducts and solve the contaminant issues in some ways.

Through relevant technical means and Meta-analysis, this project will focus on the microorganisms, which could survive in the environmental conditions of coal tar or other petroleum products contaminated media, as well as their corresponding gene sequences. By conducting the integration, comparison and analysis of these microorganisms, making the evaluation of the influences on the microbial communities caused by contaminants, as well as the judgement on the feasibility of the biodegradation of PAHs and other pollutants.

Key Words: meta-analysis, microbial diversity, coal tar, PAH, biodegradation.

Contents

1	Introduction.....	1
1.1	Background.....	1
1.2	Coal tar.....	1
1.3	Polycyclic aromatic hydrocarbons (PAHs).....	2
1.4	Biodegradation of PAHs and other contaminants.....	3
1.5	Related research on the bioremediation of coal tar contaminated soils.....	4
2	Methods.....	8
2.1	Initial preparation works on sample data.....	8
2.1.1	The selection of sample data.....	8
2.1.2	The acquisition and storage of data.....	9
2.2	The primary processing of sample data.....	9
2.2.1	QIIME2 workflow.....	9
2.2.2	DADA2 workflow.....	9
2.2.3	Deblur workflow.....	10
2.2.4	The comparison and selection between DADA2 and Deblur workflow.....	11
2.3	The further collecting and sorting on related information.....	11
2.4	The meta-analysis process and the visualization of the results.....	12
3	Results.....	14
3.1	Results of alpha diversity analysis.....	14
3.1.1	Results of alpha diversity analysis when the variable is “environmental material”.....	15
3.1.2	Results of alpha diversity analysis when the variable is “PAH concentration”.....	16
3.1.3	Results of alpha diversity analysis when the variable is “historical PAH contamination”.....	17
3.2	Results of beta diversity analysis.....	17
3.3	Results of environmental filtering analysis.....	21
3.4	Results of taxa bars plotting.....	23
3.4.1	Results of taxa bars with the variable as “environmental material”.....	24
3.4.2	Results of taxa bars with the variable as “PAH concentration”.....	26
3.4.3	Results of taxa bars with the variable as “historical PAH contamination”.....	28
4	Discussion.....	30
5	Conclusion.....	32
	Acknowledgements.....	33
	References.....	34

1 Introduction

1.1 Background

Manufactured gas, which is a kind of fuel gas and has been utilized as an energy source for lighting and heating since the early 19th century (Hatheway, 2012). This kind of fuel was generally generated from coal, coke and oil before the 1950s' at the establishment, which could be called manufactured gas plants (MGPs) (REMIEDIATING TAR-CONTAMINATED SOILS AT MANUFACTURED GAS PLANT SITES, 1994).

During the first 100 years of the development, the manufactured gas industry got rapidly thrived, particularly within the United States and Europe. It is recorded that there was at least one manufactured gas plant constructed in every Britain town or city by the mid of 19th century (McGregor et al., 2012).

However, after the mid of 20th century, the manufactured gas industry declined dramatically, which results from the utilization of natural gas and also leads to the elimination the facilities that related to gas production (Brown et al., 2006), it is reported that there were only 4 gas plants remained in the UK (Hatheway, 2012). After the structures that built above and below the ground were razed and buried respectively, massive manufactured gas plants were abandoned. It is estimated that there are more than 3000 former manufactured gas plants (FMGP) in the United Kingdom alone (McGregor et al., 2012).

The substances in manufactured gas plant waste, which have drawn attentions from environmental fields mainly include volatile organic compounds (VOCs), polycyclic aromatic hydrocarbons (PAHs), inorganic sulfur and nitrogen compounds, cyanides and heavy metals as well (Brown et al., 2006). Some of these matters above have been identified as the main pollutant that affecting human health (Collins, Williams and McIntosh, 2001). Combined with the quantity of former manufactured gas plants exist not only in the United Kingdom but also around the whole world, it has posed an extremely heavy environmental burden to the natural ecology. To solve the contamination issues that related to former manufactured gas plants has gradually become a hotspot in the relevant research fields in recent years (McGregor, 2012).

1.2 Coal tar

Coal tar is a kind of dense non-aqueous phase liquid, which is a main byproduct of manufactured gas process. it is an extremely heterogeneous material and could contain an estimated ten thousand chemicals, but only about 50% of these compounds have been identified (Franck, 1963). The components in coal tar mainly include polycyclic aromatic hydrocarbons (including 4-rings, 5-rings, 6-rings and 7-rings), methylated and polymethylated derivatives, mono- and polyhydroxylated derivatives, and heterocyclic

compounds as well (Humans, 2012). Others include benzene, toluene, xylenes, cumenes, coumarone, indene, benzofuran, naphthalene and methyl-naphthalene, acenaphthene, fluorene, phenol, cresols, pyridine, picolines, phenanthrene, carbazole, quinolines and fluoranthene (Wexler, 2014).

Coal tar is widely utilized in several fields, especially for the medical domain. It is used in medicated shampoo, soap and ointment as it could demonstrate antifungal, anti-inflammatory, anti-itch and antiparasitic properties; additionally, it could also be applied as a treatment for some common diseases or symptoms, such as psoriasis. Moreover, it had also been used in the fields of construction and industry for sealing roads and firing boilers (Williams et al., 2014).

However, as it contains a variety of detrimental substances (many of constituents in coal tar have been regarded as carcinogens, such as PAH and Benzene), coal tar is also recognized as a kind of industrial pollutant, which is urgent to be disposed (McGregor, 2012). These constituents in coal tar could contaminate the soil, surface water, groundwater and air, which could cause a huge impact on the local environmental ecology (Johnsen and Karlson, 2007).

In the meantime, as a main byproduct of the manufactured gas process, the chemical composition of coal tar could also vary as a function, which is affected by several factors, such as raw materials, plant operating conditions and even the weathering after it has been released to the environment (Brown et al., 2006). This also brings greater challenges to environmental-protection-related workers and researchers.

1.3 Polycyclic aromatic hydrocarbons (PAHs)

Polycyclic aromatic hydrocarbons (PAHs) are a kind of hydrocarbons, which are composed of multiple aromatic rings. These chemicals could be usually found in coal and tar deposits, and it has been reported that the producing of PAHs could result from the incomplete combustion of organic matters (Abdel-Shafy and Mansour, 2016).

PAHs could be considered as the one of organic compounds that has been studied most during the recent two decades (Wu, Guo, Wu and Chen, 2020). As a kind of important environmental contaminants, PAHs have been mostly identified as a kind of single contaminants, which could occur in various environmental conditions all over the world (Thavamani, Megharaj and Naidu, 2011). Due to the fact that it could be commonly found around the sites, where have been contaminated by petroleum or tar, PAHs are often associated with and pointed by the research about tar- or petroleum-contaminated soil.

The persistence in the environment and the potential of causing deleterious effects on humans, animals and even microorganisms make the degradation of PAHs become a focus in related fields (Johnsen and Karlson, 2007). With the increase of number of its rings, the changes of its chemical structure and the increase of its hydrophobicity, the electrochemical stability, durability, biodegradability and carcinogenicity of PAH would also be enhanced; in the meantime, its volatility would decrease as the growth of its molecular weight (Enzminger and Ahlert, 1987). As it has been reported that PAHs have a bioaccumulation effect in many biological chains in natural environment and the

content of them in natural world is also quite horrendous (Lee, Ong, Golchin and Nelson, 2001), PAHs have been also regarded as one of the main organic pollutants, which could affect human health. As the aspect of medicine, PAHs could cause damage to human's respiratory, circulatory, nervous systems, as well as the liver and kidneys. Another property of PAHs, which makes the contamination issue more difficult to solve, is that PAHs are hydrophobic compounds with low solubility in water. This means PAHs are more likely to bind with soil or other organic matters, which could severely limit their availability to be degraded, especially for the biodegradation (Enzminger and Ahlert, 1987).

1.4 Biodegradation of PAHs and other contaminants

During the past 30 years, several technologies have been suggested in various literature for the remediation of contaminated soils, such as heating, soil washing, solvent extraction, chemical oxidation, and bioremediation (Wu, Guo, Wu and Chen, 2020). Among these treatment approaches, the bioremediation has been identified as the most concerned and researched remediation technology in recent years (Lee, Ong, Golchin and Nelson, 2001).

Bioremediation, in a broad sense, is a process, which is used to treat the contaminated media, such as water, soil and subsurface material. The basic mechanism of bioremediation is by changing the environmental conditions to stimulate the growth of microorganisms and utilize them to degrade the target contaminants (Li et al., 2005). In most situations, the methods of bioremediation could be mainly divided into two types: bio-stimulation and bioaugmentation (Wu, Guo, Wu and Chen, 2020). Bio-stimulation is a technology, which is by adding various nutrients to the media (soil) to stimulate the native microorganisms (Hatheway, 2012); while the bioaugmentation is to inoculate some other microorganisms, which could be identified as contaminant degraders into the soil, and it is usually considered as a both suitable and efficient treatment option when the native microbial populations that could act as bio-degraders are low in the contaminated soils (Wu, Guo, Wu and Chen, 2020).

Previous studies have shown the bio-accessibility of PAHs and other pollutants in coal tar contaminated soils (Li et al., 2005), as well as the fact that some specific microorganisms such as some bacteria and fungi are capable to degrade the contaminants above, which could prove the possibility of bioremediation on such a situation (Canet et al., 2001). Previous study has pointed out that under the optimum conditions of nutrient supplementation, the biodegradation of contaminants in soil is even faster and more extensive than the process of desorption, which is also under the maximum concentration gradient between the soil and bulk solution phases that sustained by the polymer adsorbent (Thavamani, Megharaj and Naidu, 2011); and there is another study, which also reports that the process of biodegradation will be initiated immediately once the specific microbial degraders are inoculated to soils, and these microorganisms could also conduct the biodegradation effectively under the optimum environmental conditions of soil (Lee, Ong, Golchin and Nelson, 2001). These results could further demonstrate a relatively high efficiency of the bioremediation of

contaminated soils. Additionally, if the process of bioremediation is conducted by native microorganisms, it could not only be more adaptive to the local environment, but also avoid the environmental issues, which are related to the release of foreign organisms. Besides, as it is safe, ecofriendly and cost-effective as well, the bioremediation is usually considered as a promising technology for the remediation of contaminated site (Li et al., 2005).

However, there are still some limitations on the bioremediation of this situation from several aspects. As the main objectives of biodegradation in contaminated soils, it is generally believed that the PAHs are poorly bio-available, and the native microorganisms in coal tar contaminated soils could be poorly accommodated or hindered in the process of biodegradation (Canet et al., 2001); while utilizing other microorganisms as exogenous bio-degraders might encounter another issue, which could reduce the efficacy of these microorganisms and caused by the existence of indigenous bacteria in soil or the inadaptation with the actual conditions of the environment that they are inoculated. Besides, there are also some factors, which could limit the growth and reproduction of the microorganisms participating in the biodegradation, and these limiting elements would further retard the process of bioremediation (Thavamani, Megharaj and Naidu, 2011). Moreover, since the mechanisms of the bioremediation of coal tar contaminated soils are not fully studied and understood, the principles and some specific methods for enhancing the effectiveness and efficiency of the bioremediation are still unclear in some ways (Bell et al., 2016).

1.5 Related research on the bioremediation of coal tar contaminated soils

According to related resources, for the studies relevant to the process of bioremediation on the contaminated soils, which is by utilizing the biodegradation of contaminants in coal tar or other pollutants in recent years, their research contents mainly include: the species of the microorganisms that participate in the process of biodegradation and bioremediation; the procedures and related mechanisms of the biodegradation of contaminants (such as PAHs); the specific methods of enhancing the efficiency of biodegradation and bioremediation.

For the aspect of the categories of microorganisms, several previous research have shown that under the conditions of a long-term PAH contamination, some kinds of indigenous of microorganisms, which have adapted with this situation, could utilize the PAHs that are bio-available as their sole carbon or energy source, this leads to an in-situ degradation of PAHs (Bell et al., 2016). According to these findings, some microbial populations, which are PAH-degradative have been isolated and characterized by culture-dependent methods; these microbial populations are usually affiliated with some related taxonomic groups including *Sphingomonas*, *Burkholderia*, *Pseudomonas*, *Rhodococcus*, and *Mycobacterium* (Neethu et al., 2019). Besides, it is also mentioned in other articles that some species of microorganisms including *Mycobacterium*, *Bacillus*, *Rhodococcus*, *Pseudoxanthomonas*, and *Microbacterium*

had been regarded as the decomposers of PAHs (Brown et al., 2015). In the meantime, according to another study, which is related to the biodegradation of PAHs in Lake Erie, the researchers pointed out that a kind of *Mycobacterium*, which could participate in the pyrene-degrading process, had a relatively broad geographical distribution and might also play an essential role in the attenuation and cycling of PAHs under the natural conditions (Sperfeld, Rauschenbach, Diekert and Studenik, 2018).

Additionally, some microbial genes such as *nidA*, *nagAc* and *nahAc* have also been mentioned and studied. Previous research reported that these genes might be functionally associated with the biodegradation of some PAHs, which had a high molecular weight (DeBruyn, Chewning and Sayler, 2007). Another study had also researched and tested the possible relationship between the biodegradation of PAHs and these genes. The results showed that there was a significantly greater abundance of *nidA* and *nagAc* at the positions with the greatest concentrations of PAHs. In addition, according to the results of this study, it was significantly observed that the *nidA* and *nagAc* were positively correlated with the concentration and mineralization of some PAHs (such as pyrene), which could indicate that these genes could be identified as the biomarkers for the degradation of some specific PAHs (Dionisi et al., 2004).

For some related mechanisms about the process of bioremediation, it has been researched and proved that the biodegradation of PAHs and the bioremediation of contaminated soils are affected by several biotic and abiotic factors, such as the type of PAH, the plant exudate and the size fractions of the soil, etc (Valencia-Agami et al., 2019). These factors have been demonstrated that could influence the composition and abundance of the total or degradative microbial populations in contaminated soils (Ren, Ren, Teng and Li, 2015). Combined with these influencing factors above, it is widely believed that the PAHs, as well as other contaminants, may not always be degraded by indigenous microorganisms, especially for the situation that the soils have never experienced the contamination of PAHs or other similar pollutants, although the PAH degradation in soils without historical PAH contamination has also been reported in recent years' research (Neethu et al., 2019). In terms of the problems and cases, which are related to the soils without historical PAH contamination, a study, which has integrated previous relevant research, has summarized several possibilities that may occur when unpolluted soil suffers from the influence of PAH or other pollutants. On the one hand, under this situation, some microorganisms could rapidly adapt to the contamination and utilize the PAHs as their carbon or energy sources, which could result in the degradation of these pollutants; on the other hand, in contrast, the PAHs, as well as other pollutants, may pose some toxic effects on the microorganisms, which do not participate in the process of biodegradation, then leading to the loss of microbial diversity. In the meantime, there are still some populations, which may not suffer from the toxic effects, but also not degrade the contaminants in soil (Ren, Ren, Teng and Li, 2015).

Some other studies have focused and researched on the resistance effects on the biodegradation and bioremediation caused by some characters of PAHs. These effects could be mainly divided into two aspects: one is that PAH could limit their bio-

availability to microorganisms by binding themselves with organic matters or soil as PAHs are hydrophobic compounds, which have a relatively low solubility in water (Júlio et al., 2019); the other effect, which could also limit the bio-availability of PAHs is the aging effect, it has been explained as PAHs could diffuse into the micropores of the soils, which makes them unavailable for the microorganisms to generate the biodegradation anymore (Tatariw et al., 2018), and previous studies have also reported a negative relationship between the extent of mineralization and the biodegradation of PAHs with the contact time between the pollutants and soil (Sperfeld, Rauschenbach, Diekert and Studenik, 2018).

Besides, under the experimental conditions, it has been observed that with the increasing of the dosage of PAHs, the microbial richness and diversity were reduced, and the community structure of microorganisms living in soil was also changed (Valencia-Agami et al., 2019). According to the results of related research, it is also suggested that the poor ability of biodegradation of PAHs could be associated with the stability or the dramatic decrease of the abundance of the dioxygenase gene (*nidA*) in some ways (DeBruyn, Chewning and Sayler, 2007). However, the effects of PAHs and other contaminants on microbial communities may not be entirely negative. Previous studies have pointed out that the contaminants could facilitate the enrichment of some groups of microorganisms, which could participate in the biodegradation and bioremediation in that area, and this could also be doubtless beneficial to the bioremediation process of contaminated soils (Dionisi et al., 2004).

For the methods of enhancing the efficiency of the biodegradation of contaminants and the bioremediation of contaminated soils, massive studies have been done in recent years and their principles could be mainly divided into two aspects, which are to activate the microorganisms that acting as bio-degraders in soil to make them perform effectively; and to reduce the poisonousness of the contaminated soil by diluting the concentration of PAHs or other pollutants, mainly including the approaches of land treatment and composting.

According to previous studies, surfactants and organic solvents have been utilized to improve the bio-availability of PAHs (Jung et al., 2016). Researchers have found that adding surfactants to the contaminated soils could lead to positive, negative or even no effects on the biodegradation of PAHs (Brown et al., 2015). In the meantime, these surfactants could adsorb on soil particles, which could result in a long time required to completely flush them out of the subsurface of the soil (Bae et al., 2018). For the usage of organic solvents, several kinds of reagents have been used such as acetone, ethanol, squalene and paraffin oil as well (Thavamani, Megharaj and Naidu, 2011). Most of results in these studies have shown that the addition of organic solvents could expedite the degradation of contaminants. Additionally, a previous study, which focuses on the facilitation of inorganic nutrients on the indigenous microorganisms to degrade the pollutants in coal tar contaminated soils, has reported that the biodegradation could be enhanced under mixed-oxygen/denitrifying conditions; and the researchers of this study also suggested that massive macronutrients such as nitrogen and phosphorus were essential to the process of biodegradation and bioremediation, but the related

requirements for the nutrient supplements (such as whether or which kind of nutrients are required) are still unclear results from the insufficient research in this field (Jung et al., 2016). Meanwhile, some studies also pointed out that the research, which was related to the stimulation of native bio-degraders should base on the specific sites and environmental conditions (Tatariw et al., 2018).

Previous studies have reported that inoculating the PAH degraders into the contaminated soils directly could not be considered as an effective way for the bioremediation, since the results have shown that these bio-degraders did not cause significant influences on the structure of related microbial community in soils (Bae et al., 2018). This could also further demonstrate the detrimental effects of the toxicity of contaminated soils on the degrading microorganisms. Among these methods, which could decrease the poisonousness of the contaminated soil, composting has been particularly researched by many studies (Thavamani, Megharaj and Naidu, 2011). Compost, which could also be utilized as a source of both co-substrates and nutrients, has the effects of ventilation, heat preservation and improving the water retention in soil as well. The bioremediation with composting could be applied with the dual purpose of both bio-stimulation and soil fertilization (Wu, Guo, Wu and Chen, 2020). The compost could supply nutrients for native microorganisms in contaminated soil, which has been suggested by several studies, and activate them to degrade the target contaminants more effectively. As the interest on the bioremediation of oil-contaminated soils with composting was growing in recent years, the bioremediation combining composting and bioaugmentation has been regarded as an emerging in-site bioprocess (Thavamani, Megharaj and Naidu, 2011). A recent study has investigated and compared the effects of “natural attenuation”, “bioaugmentation only”, “composting only” and “bioaugmentation with composting” treatments on the biodegradation of PAHs, as well as the change on microbial community in the process of the bioremediation of oil-contaminated soils. The results showed that both “composting only” and “bioaugmentation with composting” could change the structures of microbial community in soil and enhance the microbial biodiversity as well; while the “natural attenuation” and “bioaugmentation only” did not lead to noticeable influences on that. This is mainly because that some kinds of microorganisms, which are affiliated with the compost such as *Azomonas*, *Luteimonas* and *pseudosphingobacterium* could be able to survive and then become the dominant microorganisms in contaminated soils, which could explain the mechanism of utilizing the compost to enhance the efficiency of biodegradation of pollutants. In the meantime, the authors have also suggested that the “bioaugmentation with composting” could be considered as the most effective treatment for PAH removal (Wu, Guo, Wu and Chen, 2020), which could also be a constructive opinion for the future research.

2 Methods

The research process of this project could be mainly divided into several parts, which has been explained below:

- 1) The selection, acquisition and storage of required sample data from related studies;
- 2) The primary processing of research data by utilizing relevant software and algorithms;
- 3) The further screening, collecting and sorting on information that related to data by utilizing relevant software and database;
- 4) The analysis and visualization of sample data by utilizing related software and algorithms.

All related algorithms and pipelines were provided by Dr Umer Zeeshan Ijaz (<http://userweb.eng.gla.ac.uk/umer.ijaz/>) at the University of Glasgow, as well as some related tutorials (<http://www.tinyurl.com/JCBioinformatics3>).

2.1 Initial preparation works on sample data

2.1.1 The selection of sample data

Due to the requirements of meta-analysis in this project on research samples and their corresponding gene sequences, several relevant keywords were used during the process of the selection of related studies. These keywords included: “coal tar / oil spill / crude oil”, which were used to restrict the topic of research; “Illumina MiSeq”, which was used to restrict the operation platform for gene sequencing works; “V3-V4 / V3/V4”, which were used to restrict the target region of gene sequence.

By adding the keywords above into related academic search engines (Google Scholar: <https://scholar.google.com/> and the online library of the University of Glasgow: <https://www.gla.ac.uk/myglasgow/library/>), 15 relevant studies were initially selected. Then, by consulting the relevant contents in these literatures, and using the NCBI accession numbers that given in articles, comparing the sample data of corresponding bio-project and the information related to their gene sequences as well with the relevant requirements of this project on these aspects. Most of selected studies were eliminated at this stage results from the reasons such as the data related to gene sequences were unavailable or could not be downloaded; the gene sequences provided did not meet relevant requirements; or the sizes of data provided were too large, which may lead to an overlong processing time and relatively unstable results during the subsequent process of analysis. Three studies, including a study led by Dr Caroline Gauchotte-Lindsay (<https://www.gla.ac.uk/schools/engineering/staff/carolinegauchotte-lindsay/>), whose sample data and related information have not been uploaded to NCBI database yet, were selected finally.

2.1.2 The acquisition and storage of data

The data of one study (Gauchotte-Lindsay, Aspray, Knapp and Ijaz, 2019) were provided directly by Dr Caroline Gauchotte-Lindsay, and the data of other two studies were downloaded from NCBI database. All downloaded data were stored in the remote cluster, which was also provided by Dr Umer Zeeshan Ijaz. The procedures above were all operated by related command lines under the condition of Linux operating system, which was set on MobaXterm software (<https://mobaxterm.mobatek.net/>) platform. During this process, the relevant sample data files downloaded from database, which contain the information about gene sequences were saved respectively in folders corresponding to their NCBI accession numbers.

Study	Accession Number	Number of Samples
Sonia S. Valencia-Agami_2019	PRJNA560042	8
Terrence H. Bell_2016	PRJNA317648	114
Caroline Gauchotte-Lindsay_2019	Not Available	40

2.2 The primary processing of sample data

After the data had been stored and organized by the methods described above, three algorithms were applied for the following analysis process: the QIIME2 workflow, the DADA2 workflow and the Deblur workflow.

2.2.1 QIIME2 workflow

QIIME2 (<https://qiime2.org/>) is a both extensible and decentralized microbiome analysis package with a focus on data and analysis transparency, which is widely used for amplicons analysis and could enable the researchers to start the analysis with raw DNA sequence data and finish with publication-quality figures and statistical results. As a fairly new workflow, the QIIME2 is equipped with some key features, which could be divided into several aspects:

- Integrated and automatic tracking of data provenance
- Semantic type system
- Plugin system for extending microbiome analysis functionality
- Support for multiple types of user interfaces (e.g. API, command line, graphical)

(What is QIIME 2? — QIIME 2 2020.6.0 documentation, 2020)

By utilizing QIIME2 workflow, both OTUs (under 97% or any other threshold) and ASVs could be generated in the latter cases without providing specific threshold, which are quite useful for the functional gene amplifications.

2.2.2 DADA2 workflow

DADA2 (<https://benjjneb.github.io/dada2/index.html>), which is short for the Divisive Amplicon Denoising Algorithm 2, is utilized to produce the abundance table by creating the Amplicon Sequence Variants (ASVs), which has been considered as a higher-

resolution similar of the OTUs table that could record the times of every exact amplicon sequence variant was observed in each sample. Generally speaking, there are some errors, which are brought into the sequencing data during the process of amplicon sequencing, and would complicate the interpretation of the related results seriously. In this situation, the DADA2 workflow conducts a novel algorithm, which could model the errors that generated during the amplicon sequencing, then utilizes this error model to further obtain the “true” sample composition. Due to the improvement on the simulation of errors, the DADA2 is identified as a both more sensitive and more specific workflow than the traditional OTU methods. Additionally, with the advantages such as it is reference free and could be applicable to any genetic locus, the DADA2 workflow has been widely used in related research fields and gradually replaced the traditional analysis workflow in some ways (index.utf8.md, 2020).

In this project, the process that utilized the DADA2 workflow mainly contained several steps, which have been described below:

- Activating the DADA2 algorithm under the Linux operating environment (on MobaXterm software);
- Generating the phylogenetic tree for the ASVs;
- Exporting all files, which were generated by QIIME2 workflow in the former process;
- Attaching the abundance table of ASVs with their corresponding taxonomy to produce the biom-file;
- Using the Picrust2 algorithm to conduct the functional analysis, in this step, the EC metagenome predictions, the KO metagenome predictions and the MetaCyc pathway abundance predictions were output and the number of ASVs was also shown in the output;
- Exporting relevant output files as biom-files and then as TSV files.

2.2.3 Deblur workflow

The Deblur workflow (<https://github.com/biocore/deblur>), which is also a frequently-used analysis algorithm in recent studies on related fields and differs from the DADA2 workflow, utilizes a sample by sample analysis method that could reduce both the memory requirements and the related computational demands (Nearing, Douglas, Comeau and Langille, 2018). This algorithm, which is based on the upper error rate bound along with both the mean read error rate and a constant probability of the insertion-deletions, would remove the predicted error derived reads that from the neighboring sequences (Prodan et al., 2020). This basic mechanism of the algorithm could also be explained in another way, which is the Deblur workflow compares the Hamming distances between sequence to sequence to an upper-bound error profile that combined with a relatively greedy algorithm. During this process, all sequences are sorted by abundance, then the number of predicted error derived reads will be eliminated from the counts of neighboring reads, which is also based on the Hamming distance between read to read, and also in this process, any sequence whose abundance has dropped to zero would be removed.

The main process of the Deblur workflow, which is shown by operating steps and

relevant command lines, is relatively similar to the DADA2 workflow. However, for the number of generated ASVs, due to the differences on the mechanism of analysis between these two algorithms, the disparity of this output may exist (Nearing, Douglas, Comeau and Langille, 2018).

2.2.4 The comparison and selection between DADA2 and Deblur workflow

In this project, the Deblur workflow acted as a substitution of the DADA2 workflow since the latter has shown a better sensitivity and resolution compared with the former according to related research (Prodan et al., 2020). However, in some situation, the DADA2 workflow might not be suitable and could not work properly as well due to some particular reasons such as the target region is not applicable.

In this process, both the DADA2 workflow and the Deblur were conducted on the selected sample data. The output of DADA2 workflow showed that there were 9462 ASVs generated by the algorithm, while the result of 8789 ASVs was shown in the output of Deblur workflow. However, due to the situation that in the following process of meta-analysis, it was found that there were some invalid reads at some samples (which were shown as zero at the positions that indicated the number of ASVs called by algorithm in the feature table generated by the workflow), which might result from that there could be some limitations on utilizing the DADA2 workflow to process the gene sequence data as their target region were V3-V4. In the meantime, although the number of ASVs called by Deblur was less than the one with DADA2, almost all of samples were read effectively according to the number of ASVs of each sample that shown in the feature table. Mainly based on this condition, the results and the outputs of Deblur workflow were finally selected for latter analysis.

2.3 The further collecting and sorting on related information

In order to prepare for the subsequent process of meta-analysis, some detailed information, which was related to the selected samples and their corresponding gene sequence data, was also required. Some of these contents could be checked and found from NCBI database such the accession number of each bio-sample, the corresponding name/code in paper of each sample and the geographical location where each sample was extracted, etc.; while some other information could be searched from the published literature, such as the temperature and pH value of the environment, the types and concentrations of contaminants in the environment, and the concentrations of various nutrients in the environment, etc. All of these data and detailed information were stored and organized in a table file (in CSV format, which could be directly read and utilized by following process). These contents mainly include: “the accession number of bio-project”, “the name of author”, “the accession number of SRR”, “the sample’s name in paper”, “the isolation source of the sample”, “the geographical location where the sample was extracted”, “the environmental feature”, “the environmental material”, “the experimental type”, “the oxygen concentration”, “the salinity of the sample”, “the type and corresponding concentration of nutrient”, “the pH value of the sample”, “the

temperature of the sample”, “the moisture of the sample”, “the types of contaminants (mainly for PAHs)”, “the concentration of contaminants (mainly for PAHs)”, “the historical PAH contamination”, “the organism”, “the NA type”, “the target region”, “the operating platform” and “the extraction method”.

Additionally, for clarifying the purposes and objectives of meta-analysis, several topics for comparison were determined as the form of questions. These questions, as well as some corresponding brief explanation on them have been listed below:

- Question1 – Does the Environmental Material (Soil/Seawater) affect the microbial community composition of the species that found in coal tar?

Three studies have been used in the meta-analysis process of this project, two of them researched the samples extracted from the soil and one from the seawater. As there are some differences between the nutrients and the way of energy transformation in these two media, this factor (environmental material) may affect the microbial community composition of the species found in them.

- Question2 – Does the pH Value (more or less than 7) affect the microbial community composition of the species that found in coal tar?

The pH value of the environment has a vital impact on the growth and reproduction of the microorganisms that living in it. The tolerance to environmental pH value could vary with different species of microorganisms; in the meantime, the activity of biochemical reaction of one kind of microorganism could also differ with the variation on pH value. Therefore, the environmental pH value could also be a factor affecting the microbial diversity in the sample.

- Question3 – Does the Concentration of PAHs (High/Medium/Low) affect the microbial community composition of the species that found in coal tar? (where the “high”, “medium” and “low” respectively stands for a PAHs concentration of “>10 mg/g”, “0.1~10 mg/g” and “<0.1 mg/g”)

It has been reported by related research that PAHs have a certain biological toxicity. Therefore, the concentration of PAHs in the environment should cause a relatively huge impact on the organisms that living in it, and the same is true for the various microbial population.

- Question4 – Does the Historical PAHs Contamination (Experienced/Not Experienced) affect the microbial community composition of the species that found in coal tar?

Previous studies have pointed out that there are some significant differences on the microbial responses to the PAHs or similar contamination between the soil (or some other environmental media) that has experienced contamination before and the soil has not. This point may be another crucial factor, which may lead to some influences on the structure of microbial community in the contaminated media.

All of these four questions, as well as related options, were also organized in the meta-data table for a direct usage in the following analysis process.

2.4 The meta-analysis process and the visualization of the results

As the main body of the meta-analysis of this project, this part was operated on the

platform of RStudio (<https://rstudio.com/products/rstudio/>). RStudio is an integrated development environment for R language, which mainly contains a console, a syntax-highlighting editor and the tools that for plotting, history, debugging and workspace management as well (RStudio, 2020).

Based on related algorithms and packages that could be run in R condition, four aspects of diversity analysis had been conducted, which includes the Alpha Diversity analysis, Beta Diversity analysis, the Environmental Filtering analysis and the Taxa Bars plotting. Several R packages that used in the process of analysis have been listed and explained briefly below:

- phyloseq, which is an essential tool that could import, store, analyze and display the complex phylogenetic sequencing data graphically that has been clustered into the Operational Taxonomic Units (OTUs) already (phyloseq: Explore microbiome profiles using R, 2020);
- vegan, which is equipped with the most basic functions that used in the diversity analysis, community ordination and dissimilarity analysis, and could provide the tools for descriptive community ecology (CRAN - Package vegan, 2020);
- data.table, which could be identified as an improved version of data.frames, served as the standard data structure for the data storage in the R condition (Extension of 'data.frame' [R package data.table version 1.13.0], 2020);
- ggplot2, which is a system that could create graphics declaratively based on 'the Grammar of Graphics' (Create Elegant Data Visualisations Using the Grammar of Graphics [R package ggplot2 version 3.3.2], 2020);
- ape, which is a package that could provide various functions for reading, writing, manipulating, analyzing, simulating phylogenetic trees and DNA sequences, computing DNA distances, translating into AA sequences, estimating trees with the distance-based methods, and a range of methods for comparative analysis and analysis of diversification (Analyses of Phylogenetics and Evolution [R package ape version 5.4-1], 2020).

3 Results

The results of the meta-analysis in this project could be mainly divided into four parts, Alpha Diversity, Beta Diversity, Environmental Filtering and the Taxa Bars plotting, which could correspond to the types of analysis that have been conducted respectively.

3.1 Results of alpha diversity analysis

In the analysis process that utilized in this project, the results of Alpha Diversity contained five important evaluation indices, which are “Richness”, “Shannon Index”, “Pielou's Evenness”, “Fisher Alpha” and “Simpson Index”, and brief introductions of these five indices have been explained below:

1) Richness, also called as “Species Richness”, which could indicate the sum of the number of species with the abundance greater than 0 in the community. It could be regarded that with the value of richness getting larger, the abundance of species in the community becoming greater. During the process of computing, the richness index will treat all existing species (no matter dominant species or rare species in the community) with equal weight, and pay attention to the existence of species without regarding their relative abundance.

2) Shannon Index, also called as “Shannon Entropy Index” or “Shannon-Wiener Index”, which could consider both the richness and evenness of the species simultaneously. It is considered as an extension of the information theory and could reflect the uncertainty of which species that can be predicted from randomly selected individuals in a community. Under this concept, it is not difficult to find that if one or a few species in a community occupy a dominant position (compared with other species, they have a clear advantage in abundance), then the uncertainty would not be high compared with the community that has a relatively higher evenness.

3) Pielou's Evenness, also called as “Shannon's Evenness”, which could be defined as the ratio of the actual Shannon Index of the community to the largest Shannon Index that could be obtained from the community with the same species richness, and if all species have the same relative abundance in the community, the value of Pielou's Evenness would be 1.

4) Simpson Index, which also takes both the richness and evenness of the species into consideration, while it is influenced heavily by the evenness compared with the Shannon Index. It could represent the probability that two randomly selected individuals in the community do not belong to the same species, which means there is a positive correlation between the value of the Simpson Index with the species abundance in the community.

5) Fisher Alpha, which is considered as theoretically independent of the size of sample, could be calculated only with the species richness and the total number of individuals. It is another crucial index that could indicate the diversity condition in a community and has been widely used in related studies.

3.1.1 Results of alpha diversity analysis when the variable is “environmental material”

When the variable is set as the environmental material (soil/seawater), it could be found that there are significant differences in both the richness and evenness of species in the samples that extracted from different environmental media.

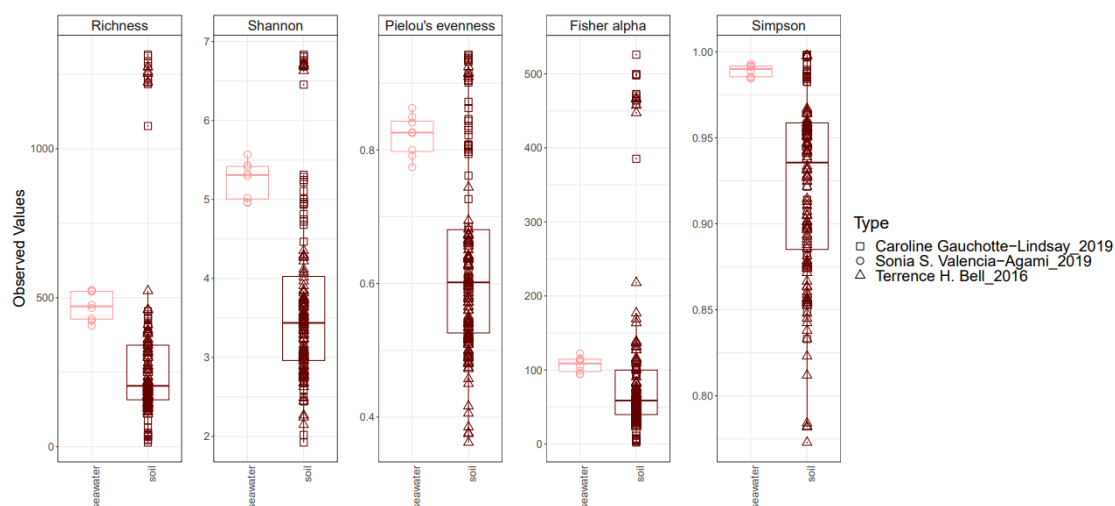


Figure 3-1 The results of alpha diversity analysis - Variable: Environmental Material

For the aspect of richness, it could be observed that the number of species that found in the samples from seawater distributed in the range of 400-550; while for the samples extracted from soil, this number mainly ranged from 50-450. Although the value of richness of a few samples extracted from the soil could reach more than 1000, the value of the richness in most soil samples still mainly concentrated in the interval of 150-300; in the meantime, the middle value of the richness of each soil sample (about 200) is also obviously lower than the number of the samples that extracted from seawater (about 450).

For the aspect of evenness, a similar situation could be observed, it could be considered that the samples extracted from seawater had a relatively higher evenness of species compared with most of samples that extracted from soil according to both Shannon Index and Pielou's Evenness. Additionally, based on the value of Simpson Index, the samples from seawater also had a relatively higher species abundance as the middle value of Simpson Index in these samples is about 0.98, compared with the samples from soil, which is about 0.94 and slightly lower than the former. However, it should be noticed that there were a few of samples that extracted from soil had an extremely high species richness and evenness as their values of Simpson Index were approaching 1.0.

3.1.2 Results of alpha diversity analysis when the variable is “PAH concentration”

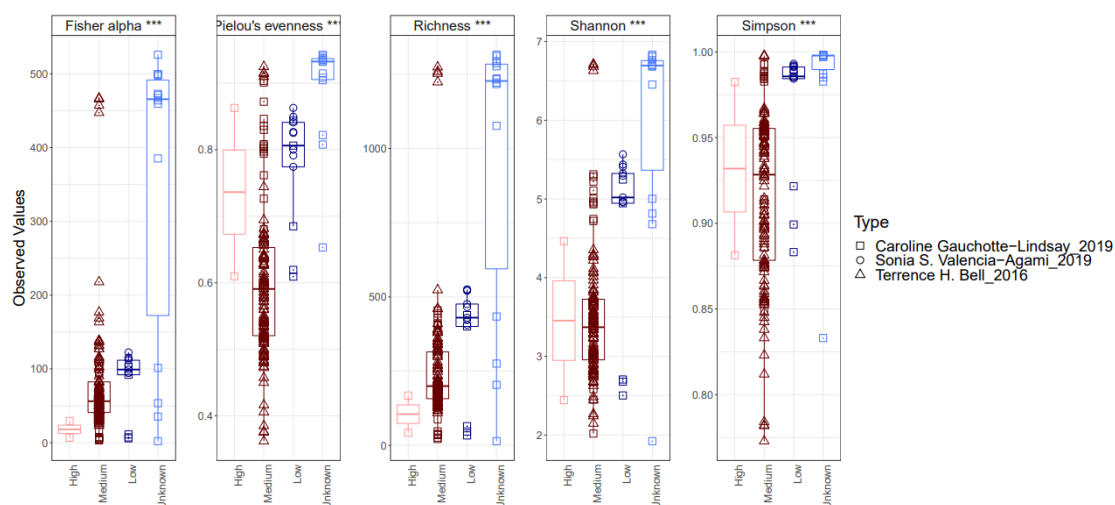


Figure 3-2 The results of alpha diversity analysis - Variable: PAH Concentration

When the variable is set as the PAH concentration (high/medium/low), different trends were observed in terms of different index. In the aspect of richness, it could be found that the richness in samples would decrease with the increase of PAH concentration. The value of richness of most of samples with low PAH concentration distributed in the range of 400-500 with a middle value of 450, which is significantly higher than the samples with medium PAH concentration (about 200) and high PAH concentration (about 100).

However, as the aspect of species evenness, according to the value of Pielou's evenness, although the species evenness in the samples with high PAH concentration are lower than the samples with low PAH concentration, the middle value of these samples (about 0.75) was still higher than the value of samples with medium PAH concentration (about 0.58).

In the meantime, when it came to Shannon Index and Simpson Index, it could be observed that the relative distributions of the samples of these three groups were extremely close. According to the meanings of these two indices, it could be inferred that the relative species abundance in the samples with low PAH concentration was much higher than the samples with high or medium PAH concentration, while the conditions in other two groups were relatively close.

3.1.3 Results of alpha diversity analysis when the variable is “historical PAH contamination”

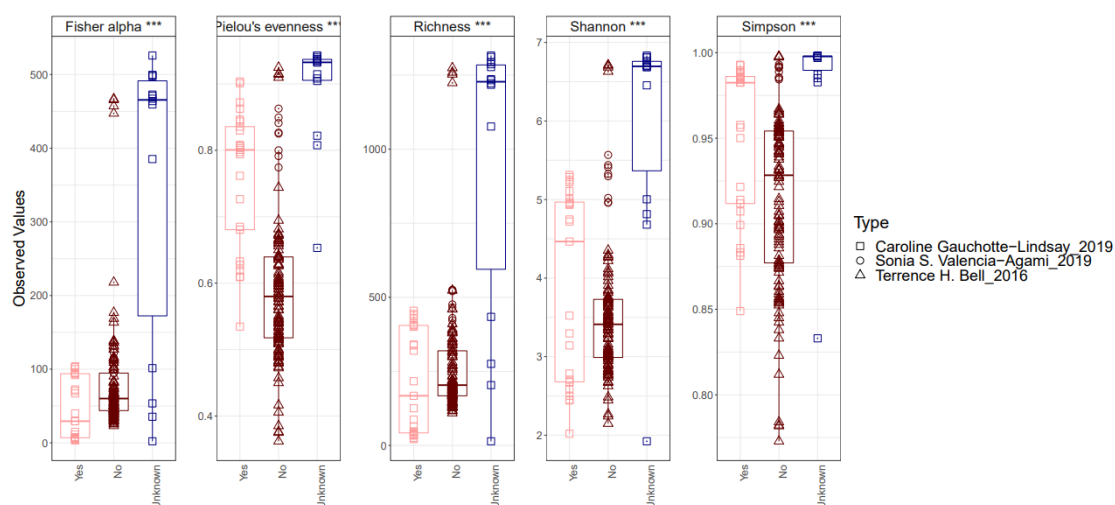


Figure 3-3 The results of alpha diversity analysis - Variable: Historical PAH Contamination

When the variable is set as the historical PAH contamination (yes/no), the situation had changed again. For the aspect of species richness, it could be found that the samples, which had experienced PAHs contamination before had a relatively lower richness (ranged from 50-450 with a middle value about 150) than the samples who had not (ranged from 120-550 with a middle value about 200). But based on other indices such as Shannon Index and Simpson Index, the situation just went opposite. It means both the species evenness and relative abundance in the samples that had experienced PAHs contamination were relatively higher than the samples had not.

Another point that should be noticed is although the historical PAHs contamination had been all experienced, the samples in this group were still distributed more scattered in terms of both richness and evenness, compared with the samples without historical PAHs contamination.

3.2 Results of beta diversity analysis

Differing from the Alpha Diversity, which focuses on the biodiversity in the samples, the Beta Diversity is utilized to compare the biodiversity between different ecosystems (which could be also considered as different samples). Beta Diversity uses the evolutionary relationship and the abundance information between the sequences in each sample to calculate the distances between samples, which could reflect whether there are significant differences in microbial community between samples.

There are several different Beta Diversity analysis measures, among these methods, the Bray-Curtis Distance, Unweighted UniFrac and Weighted UniFrac had been conducted in this project.

- Bray-Curtis Distance, whose advantage is a simple computing process, takes both the species abundance and the evenness into consideration simultaneously. However, in this type of Beta Diversity analysis, the evolutionary relationship between OTUs (or

ASVs) has not been considered;

- Unweighted UniFrac, which conducts the comparison according to their phylogenetic trees, could classify the OTUs (or ASVs) on the basis of their 16S sequence information. The Unweighted UniFrac only considers if there are any changes in the aspect of species;

- Weighted UniFrac, whose basic principles are similar to the Unweighted UniFrac, considers the changes both in the existence of species and the abundance of species.

Additionally, in the process of Beta Diversity analysis, two main analysis procedures would be usually used, which are PCA (Principal Component Analysis) and PCoA (Principal Co-ordinates Analysis), a brief introduction of them has also been shown below:

- PCA (Principal Component Analysis), which is a visualization method to study the similarity or discrepancy between data, adopts the idea of dimensionality reduction and could find the most important coordinates in the distance matrix. After sorting the complex data with a series of eigenvalues and eigenvectors, it would select the main first few eigenvalues to indicate the relationship between different samples. Based on this theory, the PCA could observe the differences between individuals or groups;

- PCoA (Principal Co-ordinates Analysis) is a dimensionality reduction method, which is similar to PCA method. The main difference between the PCoA and PCA is that the PCA bases on the analysis of the original species composition matrix, it uses the Euclidean Distance and only compares the difference on the abundance of species; while the PCoA will compute the distances between samples according to different algorithms firstly and then process the distance matrix, in order to achieve the quantitative conversion of qualitative data.

In the process of the Beta Diversity analysis in this project, the PCoA method is mainly utilized. The degree of the differences of species between samples could be reflected by the distance between the corresponding points of different samples in the figure. The greater the distance between two points, the greater the differences of species between two samples.

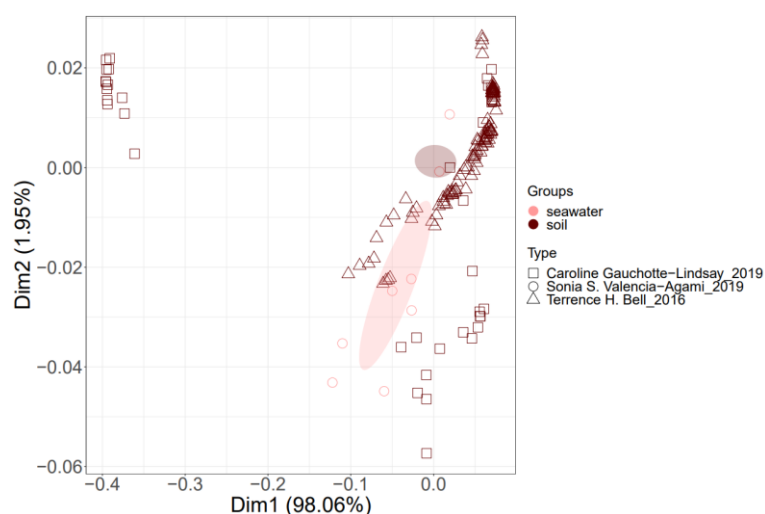


Figure 3-4 The result of beta diversity analysis - Variable: Environmental Material (Weighted UniFrac)

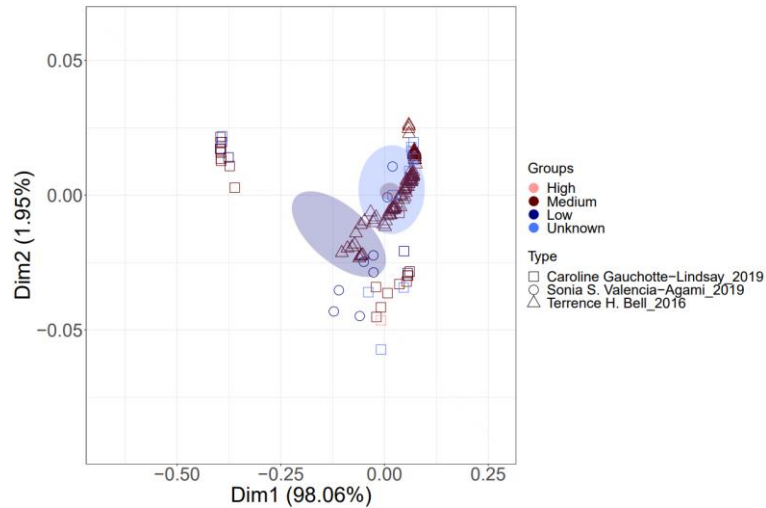


Figure 3-5 The result of beta diversity analysis - Variable: PAH Concentration (Weighted UniFrac)

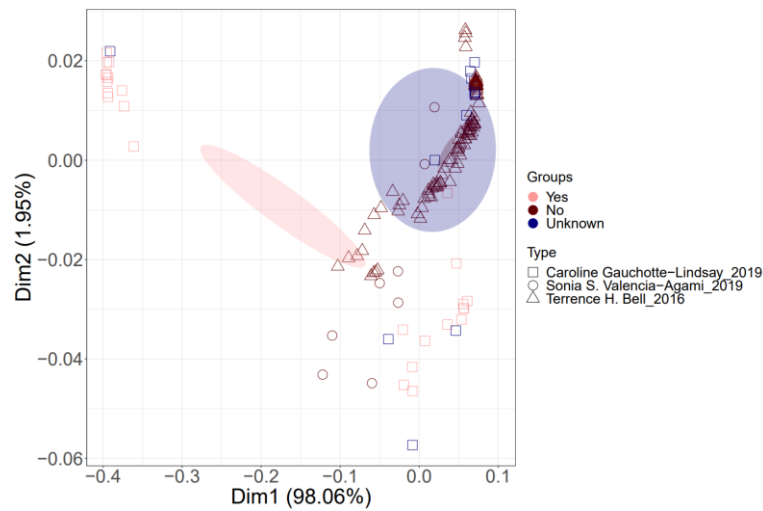


Figure 3-6 The result of beta diversity analysis - Variable: Historical Contamination (Weighted UniFrac)

According to the figures of results of Beta Diversity analysis, it could be found that although the relative positions of points (which are corresponding to the different samples that were contained in the analysis) in the figures are nearly the same (as the samples, which were included in these analysis did not change), then relative relationship between different groups still changed when the type of variable was switched.

In terms of the overall distribution, it could be observed that except for a small number of discrete samples, the distribution of most samples is relatively concentrated, which could indicate that the microbial communities in these samples have a certain similarity in some ways. However, among these figures, it is also worth noticing that the situation of distribution of samples in some areas is extremely dense, while it is relatively scattered in some other parts, which could show that there are still certain differences in the structure of the microbial communities between a large part of samples.

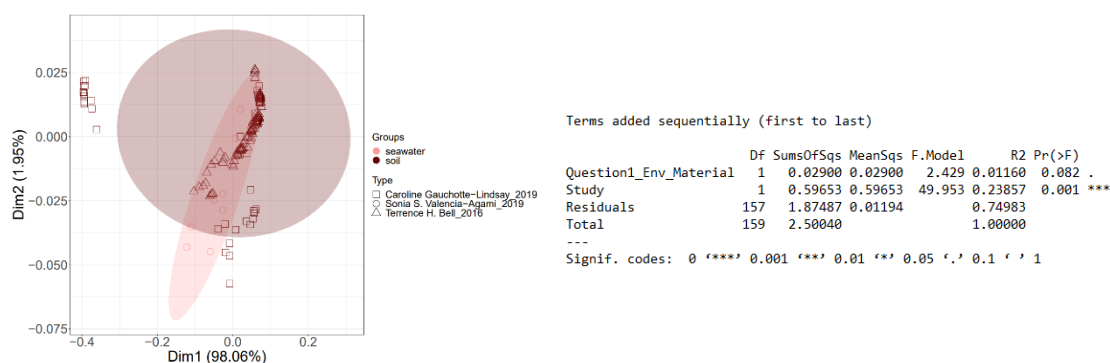


Figure 3-7 The numerical result of beta diversity analysis - Environmental Material

When considering different variables that were set in the Beta Diversity analysis, there are still some dramatic differences between these figures of results. For the analysis with the variable as “environmental material (soil/seawater)”, it could be observed that most of samples extracted from soil are densely distributed and only a limited part of samples are scattered, which could reflect that the compositions of the microbial community in the soil samples are relatively similar; while the distribution of the samples extracted from seawater is more discrete. Additionally, according to the correlation between the areas covered by the two groups of samples, it is not difficult to find that although there are certain differences between these samples, a relatively high similarity in the composition of the microbial community in the two groups of samples could be shown. This point could also be reflected by the numerical results generated by the algorithm that used for analysis, where it is shown that the P value between the two groups of samples is equal to 0.082, indicating the structures of microbial communities in the samples of two groups are relatively similar.

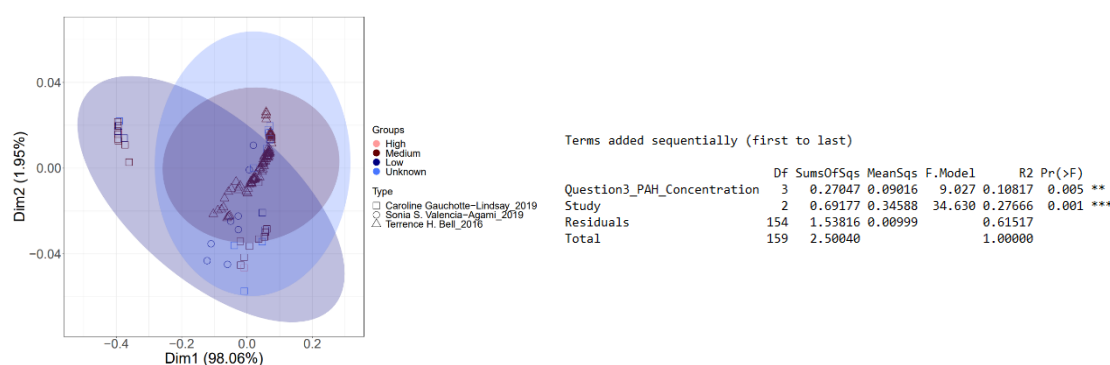


Figure 3-8 The numerical result of beta diversity analysis - PAH Concentration

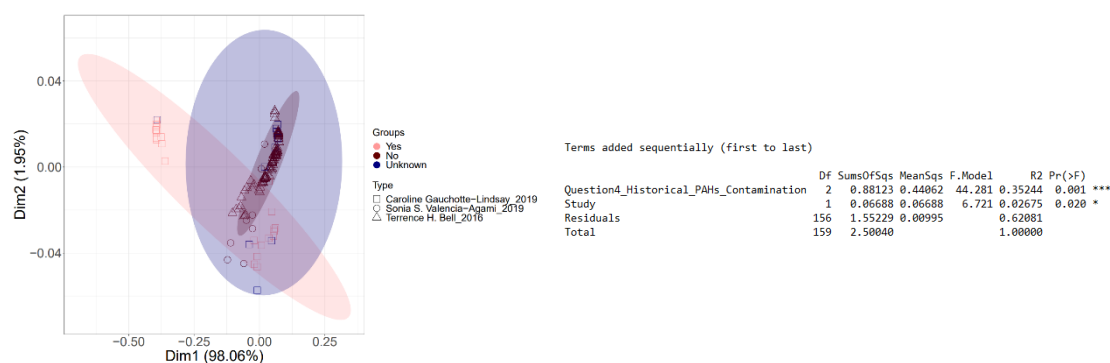


Figure 3-9 The numerical result of beta diversity analysis - Historical PAH Contamination

When the variables are switched to “PAH concentration (high/medium/low)” and “historical PAH contamination (yes/no)”, it is not difficult to find that the overlapping areas between the ellipses that corresponding to different groups of samples become smaller and smaller. This could indicate that when switching to these two variables, the differences in the composition of the microbial communities of the samples that contained in different groups become larger and larger. Similarly, this could also be reflected by the numerical results generated by analysis algorithm. When the studied variable is PAH concentration, the P value is 0.005, which could indicate a relatively large difference; and when the studied variable is historical PAH contamination, then the P value becomes 0.001, the differences in the structure of microbial communities between different groups become larger.

3.3 Results of environmental filtering analysis

The theory of environmental filtering mainly focuses on the relationship between the environmental conditions and the organisms that living in them. It suggests that not all organisms could be able to survive, grow and reproduce normally in various environmental conditions; in other words, the environment could be regarded as a selective force and eliminate the species, which are unable to adapt to the conditions at a particular location.

The influence of the environmental filtering on the microbial community could be expressed by corresponding figures. It is mainly reflected by the ratio of NRI value to NTI value of different samples.

- When the value of NRI/NTI greater than 0, it means that there is an environmental pressure on the microbial community; and when it goes to more than +2, it could indicate that the effect of environmental filtering is relatively strong;
- When the value of NRI/NTI less than 0, in contrast, it means that there is a competitive exclusion in the microbial community; similarly, when it goes to less than -2, it could be inferred that the competitive exclusion is relatively strong.

According to the figure results generated by the environmental filtering analysis, which focused on different variables, it could be found that there are significant differences between the results of the analysis of environmental filtering or environmental pressure when the studied variable switched.

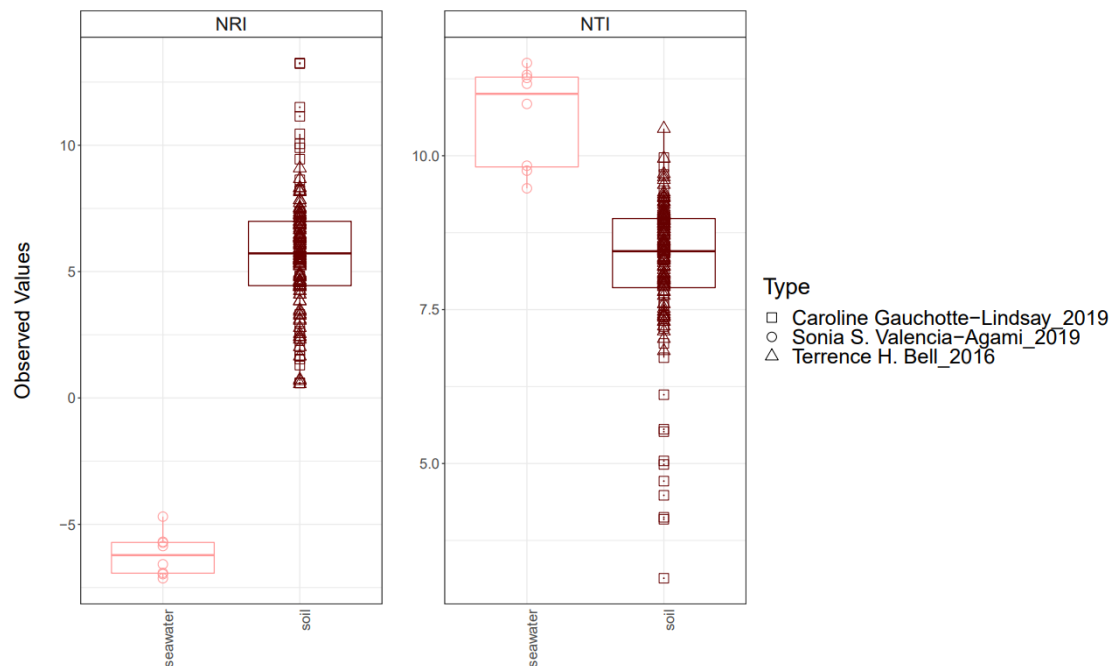


Figure 3-10 The result of environmental filtering analysis - Variable: Environmental Material

When the variable is set as “environmental material (soil/seawater)”, it could be obtained that for the samples extracted from seawater, the value of NRI/NTI ranged from -1~0, which could indicate a competitive exclusion in the microbial community in this group of samples; while for the samples extracted from soil, the values of this ratio are all greater than 0, and some samples even have a NRI/NTI with the value of more than 2, which could indicate a relatively strong effect of environmental filtering.

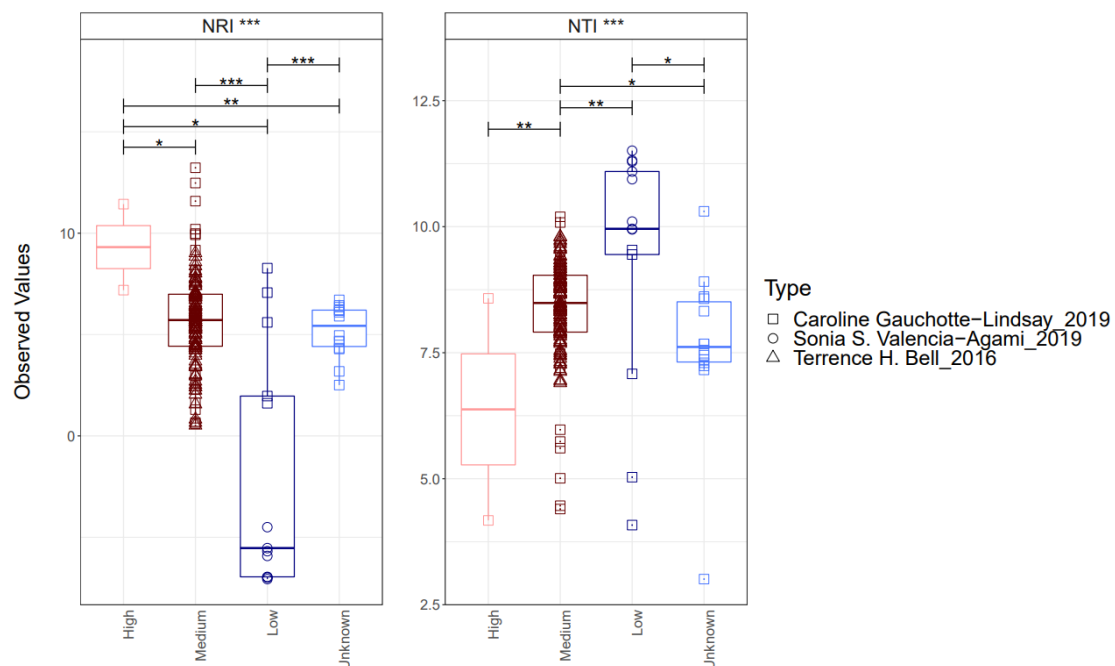


Figure 3-11 The result of environmental filtering analysis - Variable: PAH Concentration

When the studied variable switched to “PAH concentration (high/medium/low)”, it could be observed that there is a trend of the variation of the NRI/NTI value. When the PAH concentration getting higher (from “low” to “medium” then to “high”), the value of this ratio would also increase, which means with the increase of the PAH concentration in samples, the effect of the environmental filtering or environmental pressure would be stronger and stronger compared with the competitive exclusion.

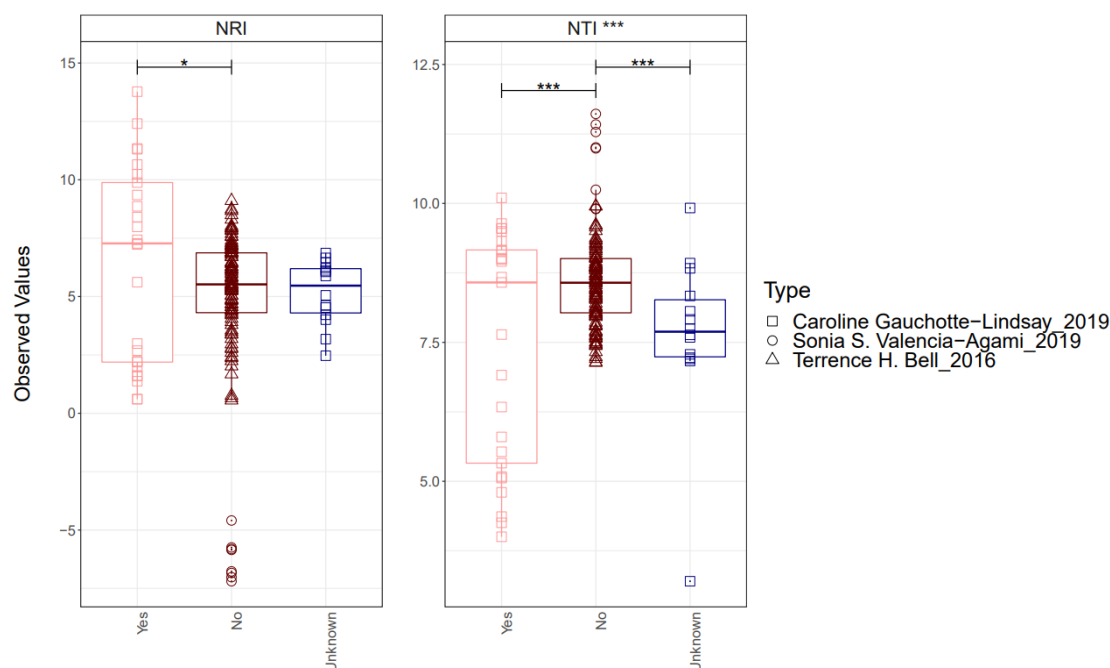


Figure 3-12 The result of environmental filtering analysis - Variable: Historical PAH Contamination

When the studied variable switched to “historical PAH contamination (yes/no)”, it could be found that the values of NRI/NTI of samples in both two groups are all greater than 0, which means there is an effect of environmental filtering in both the samples that have or have not experienced historical PAH contamination before. Besides, it could also be estimated that the values of NRI/NTI of samples that have experienced historical PAH contamination are slightly higher than the samples have not, which means compared with the samples without historical PAH contamination, the samples with it would be influenced more by the effect of environmental filtering, or bear more environmental pressure.

3.4 Results of taxa bars plotting

Taxa bars could be used to reflect the specific types and proportions of various microorganisms that discovered in the samples. According to relevant principles and methods of biological taxonomy, any organism could be classified into the levels of “Kingdom”, “Phylum”, “Class”, “Order”, “Family”, “Genus” and “Species”; and in the process of biodiversity analysis, the results would also change with the different selected level. In the Taxa Bars analysis of this project, the levels of “Class”, “Order”, “Family” and “Genus” would be mainly focused.

3.4.1 Results of taxa bars with the variable as “environmental material”

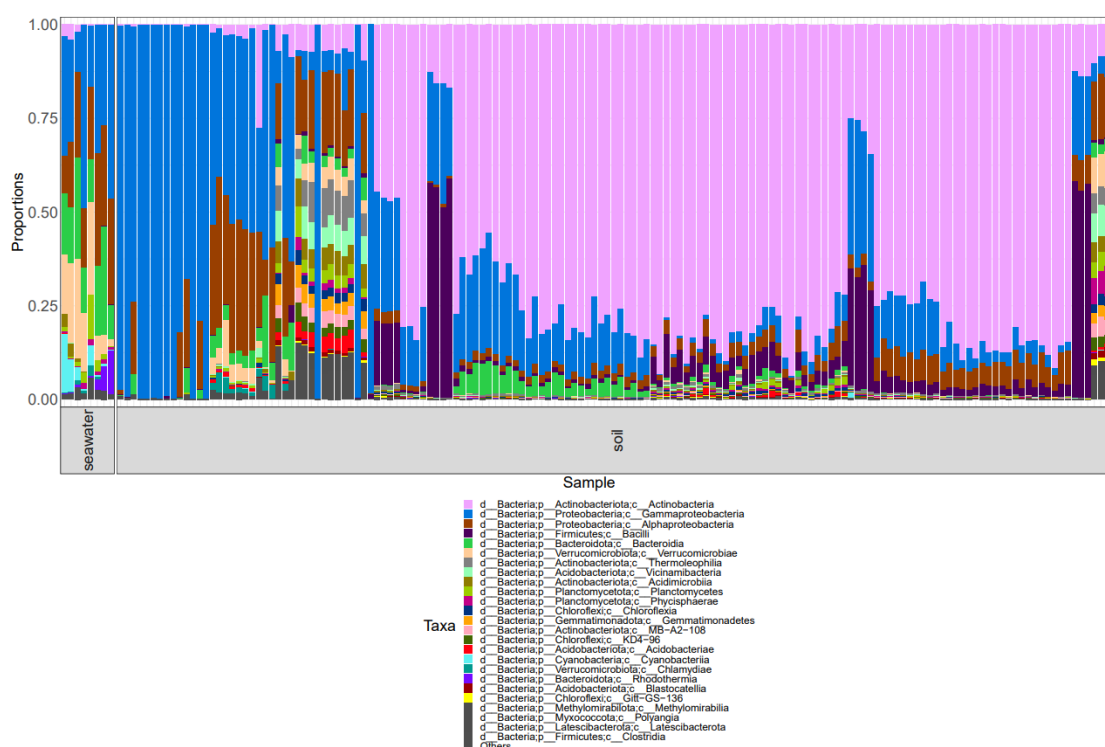


Figure 3-13 The result of taxa bars figure - Environmental Material - Class

When the variable is set as “environmental material (soil/seawater)”, it could be observed that there are significant differences on both the structure and composition of the microbial communities between the samples of two groups. For the level of class, it is obvious that the actinobacteria and gammaproteobacterial occupied a quite large proportion of the entire microbial community in most of samples; while for a small number of soil samples and almost all samples that extracted from seawater, the gamma-proteobacteria and alpha-proteobacteria are the major types of bacteria found in them. In the meantime, it could be observed that the samples extracted from seawater have a relatively higher evenness than the soil samples, and for several samples from soil, which is also noticeable, the gamma-proteobacteria accounted for nearly all proportions of the microbial community.

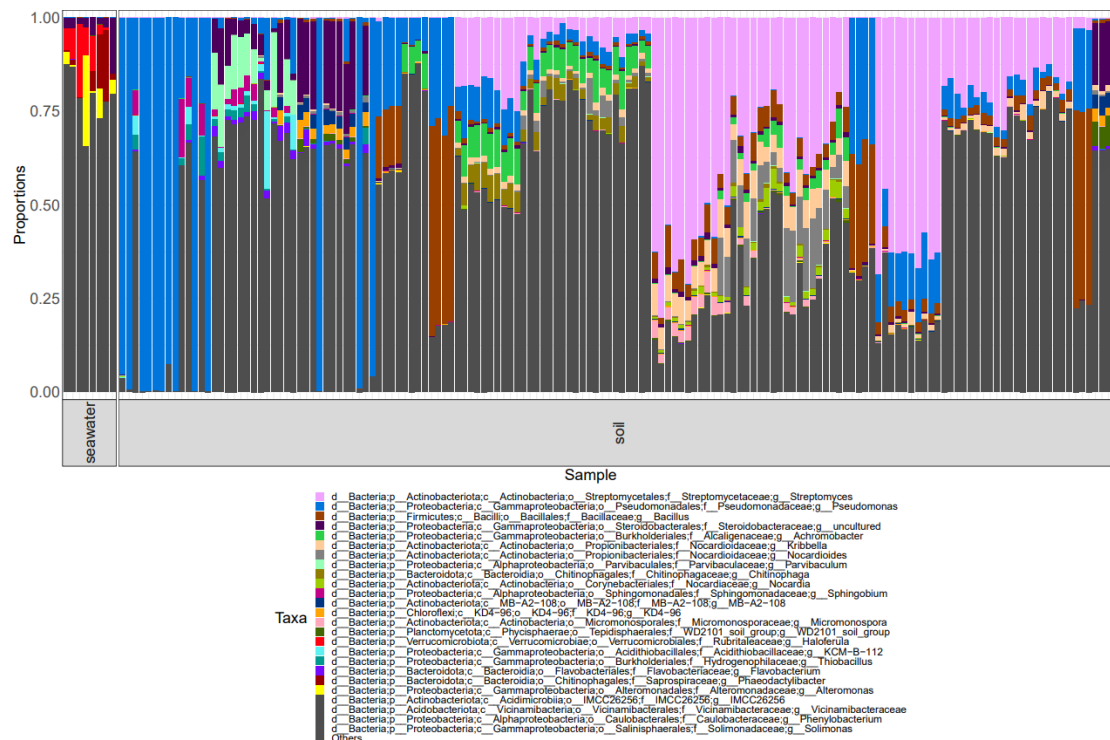


Figure 3-14 The result of taxa bars figure - Environmental Material - Genus

When the studied level is set as genus, it could be found that the analysis results of the samples extracted from seawater have a higher similarity, where the *Haloferrula*, the *Phaeodactylibacter*, the *Alteromonas* and another uncultured genus, which belong to *Rubritaleaceae* family, *Saprospiraceae* family, *Alteromonadaceae* family and *Steroidobacteraceae* family respectively, occupied the largest proportions of the microbial community. While the situations in the soil samples are more complicated, where the *Streptomyces*, the *Pseudomonas*, the *Bacillus* and the *Achromobacter*, which belong to *Streptomycetaceae* family, *Pseudomonadaceae* family, *Bacillaceae* family and *Alcaligenaceae* family are the dominant genus of microorganisms in most of soil samples; but for a small number of samples, the *Parvibaculum* that belongs to *Parvibaculaceae* family and the *Sphingobium* that belongs to *Sphingomonadaceae* family could also be considered as major genus.

3.4.2 Results of taxa bars with the variable as “PAH concentration”

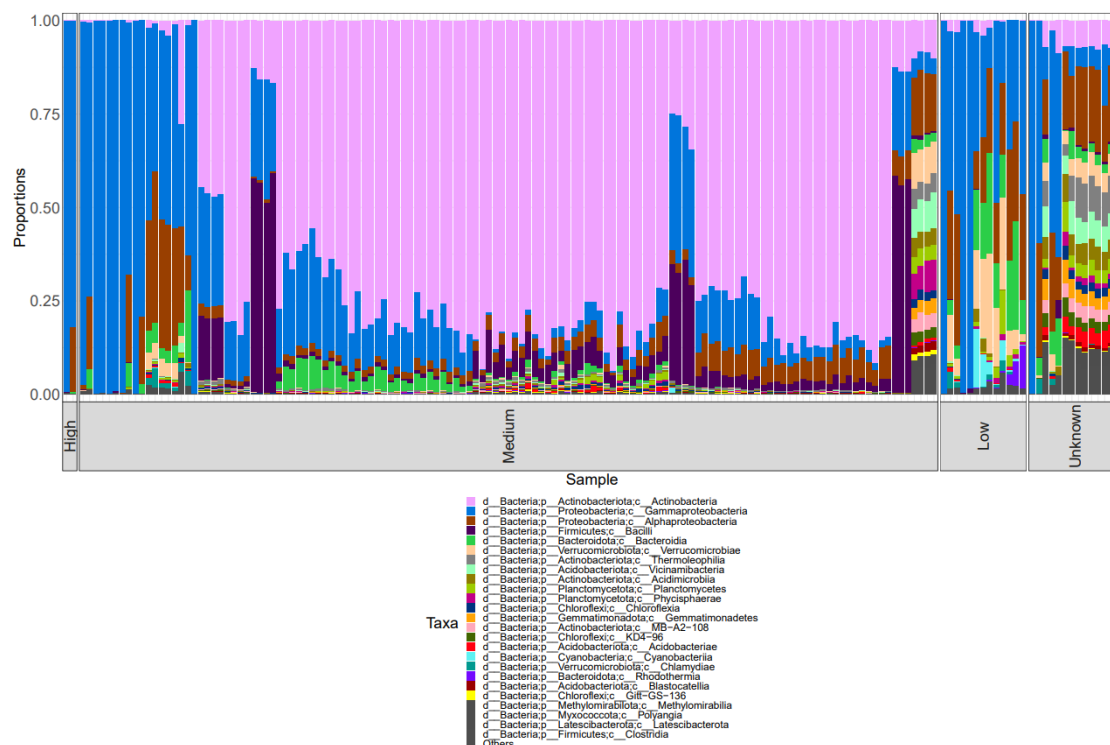


Figure 3-15 The result of taxa bars figure - PAH Concentration - Class

When the variable is set as “PAH Concentration (high/medium/low)”, at the level of class, it could be observed that the microorganisms that belong to Gamma- and Alpha-proteobacteria occupied a relatively high proportion of microbial community in almost all samples; while in a number of samples with medium PAH concentration, the microorganisms that belong to Actinobacteria could account for even more than 50% of the entire community. Additionally, it could be also found that with the increase of PAH concentration in samples, the evenness of microbial community was reduced.

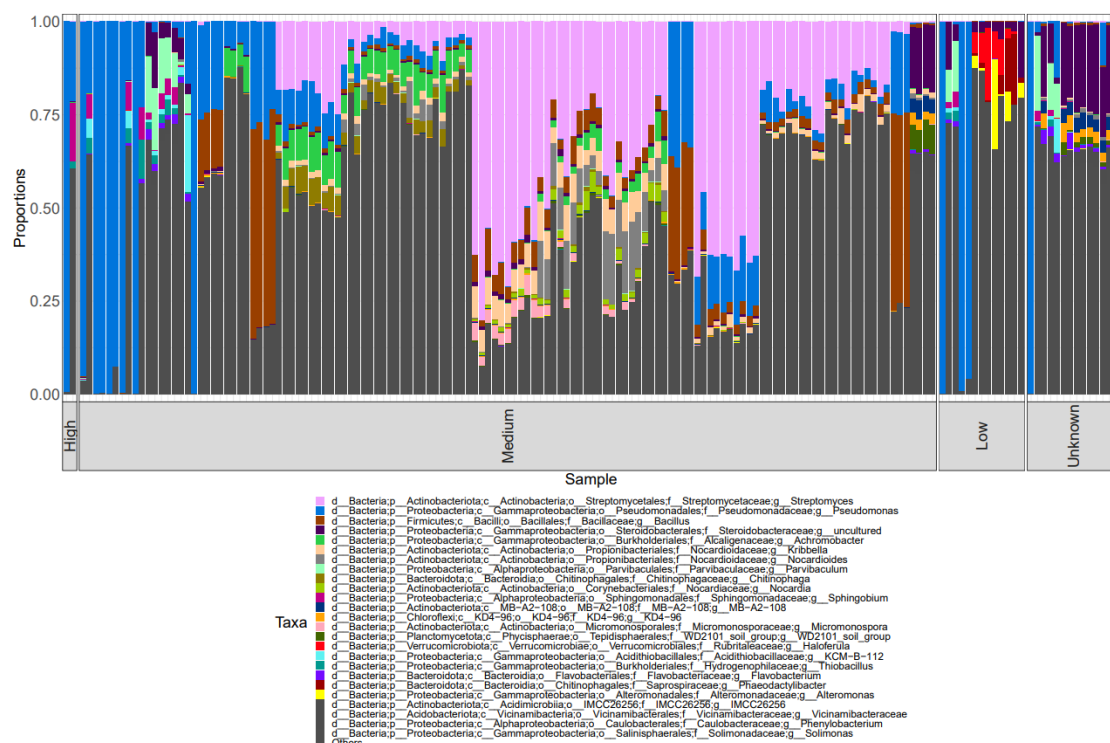


Figure 3-16 The result of taxa bars figure - PAH Concentration - Genus

At the level of genus, it is obvious that the *Pseudomonas* that belongs to Pseudomonadaceae family could be identified as the dominant genus in most of samples, as well as the *Streptomyces* that belongs to Streptomycetaceae family, but there are still great differences on the exact proportions of these major genus between different samples.

3.4.3 Results of taxa bars with the variable as “historical PAH contamination”

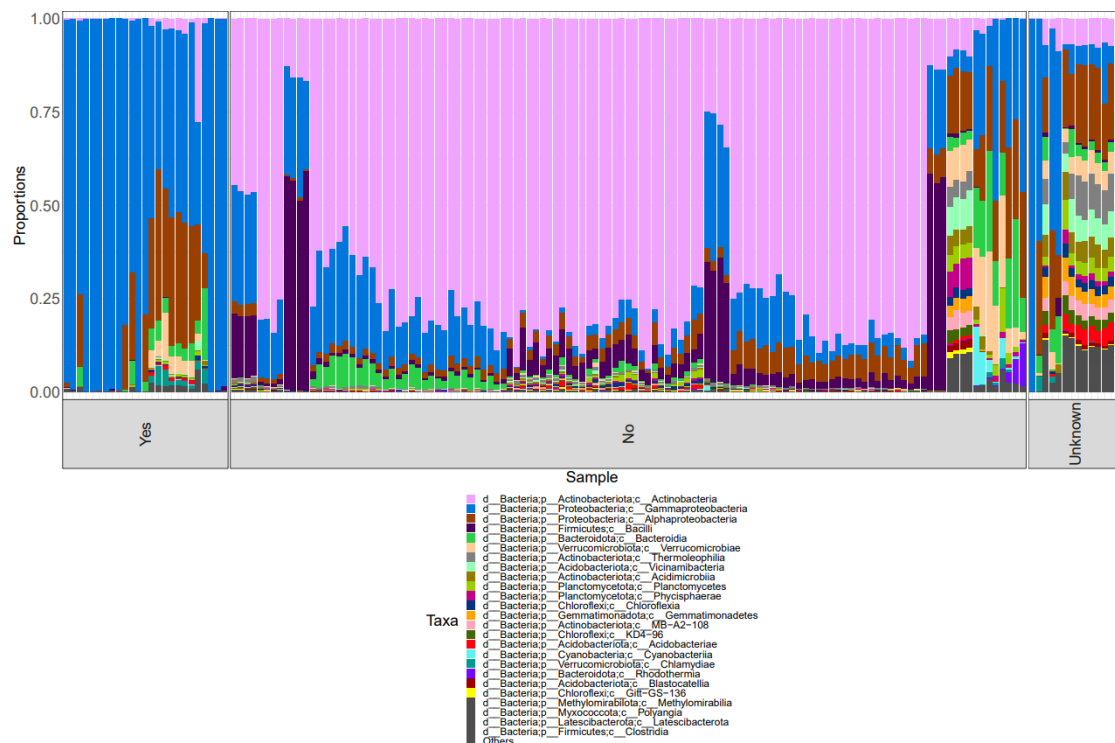


Figure 3-17 The result of taxa bars figure - Historical PAH Contamination – Class

When the variable is set as “historical PAH contamination (yes/no)”, at the level of class, it could be observed that there are obvious differences on the structure of the microbial communities between the samples with historical PAH contamination and the samples without. The gamma- and alpha- proteobacteria occupied a certain proportion of the entire community in all samples, but the specific ratios of them changed (decreased) dramatically from the samples that had experienced historical PAH contamination to the samples had not; while for the most of samples without that, it is the actinobacteria that accounted for the largest percentage of the microbial community.

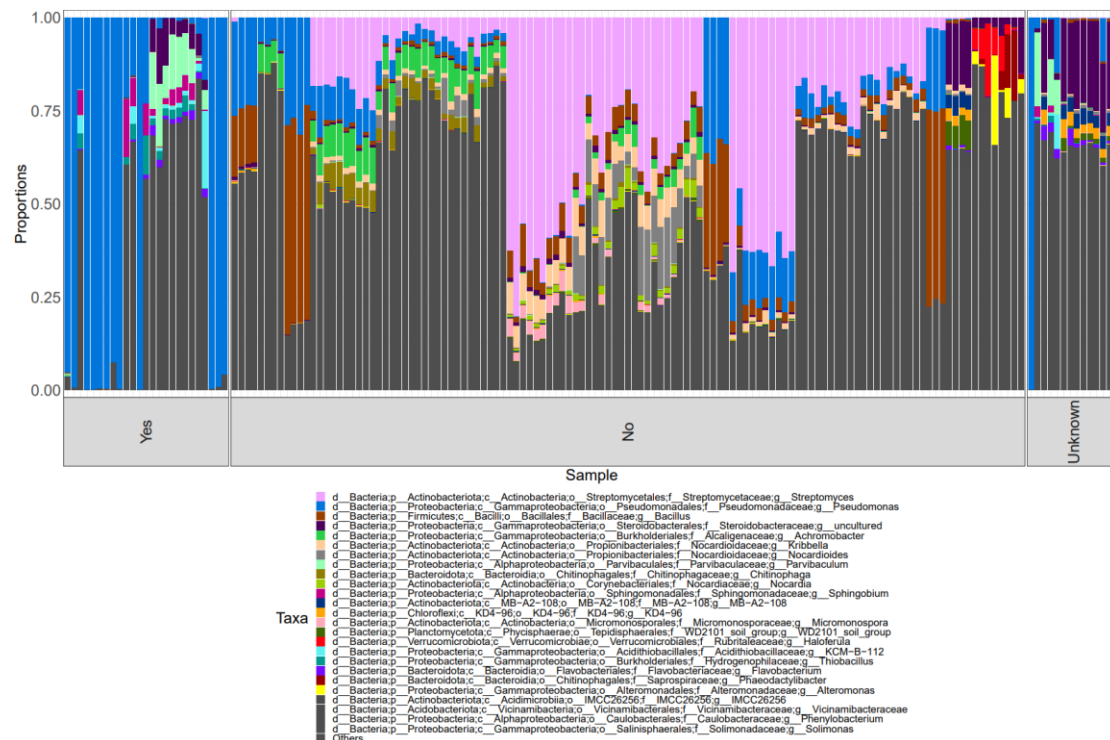


Figure 3-18 The result of taxa bars figure - Historical PAH Contamination - Genus

As the studied level switched to genus, the differences between the samples of two groups became more significant. It could be recognized that the Pseudomonas that belongs to Pseudomonadaceae family, the Parvibaculum that belongs to Parvibaculaceae family, the Sphingobium that belongs to Sphingomonadaceae family and the genus coded as “KCM-B-112” that belongs to Acidithiobacillaceae family are the dominant genus of the microbial community in the samples with historical PAH contamination. While for the samples that had not experienced contamination before, the major genus in the community changed to the Streptomyces that belongs to Streptomycetaceae family, the Bacillus that belongs to Bacillaceae family, the Achromobacter that belongs to Alcaligenaceae family and the Kribbella that belongs to Nocardioideae family.

4 Discussion

From the results of alpha diversity analysis, it could be obtained that the PAH contamination could influence both the richness and evenness of the microbial communities, especially on the richness aspect, this has also confirmed the point “PAHs have certain biological toxicity to the microorganisms living in the environment” in some ways, which had been stated by some related studies. However, from the aspect of evenness, it could be found that the samples with high PAH concentration have a slightly higher degree of evenness than the samples with medium PAH concentration. This might result from that the high concentration of PAHs eliminated a large proportion of microorganisms, which could not adapt to such a situation, while some other species that could survive then became the dominant species and maintained a relatively high evenness in the entire community. On the other hand, it could also be explained as the number of samples, which belong to the group with high PAH concentration is too small (only 2), compared with the samples, which had a medium PAH concentration (more than 100), this may cause a certain error in the process of analysis, which ultimately leads to the gap between the results obtained and the expectations. Additionally, when the analysis focused on the historical PAH contamination, similar results could be seen on the aspect of richness, this could be mainly because that some species might have been excluded under the environmental condition of a long-term effect of PAHs. While from the aspect of evenness, different results were shown, this could be explained as, for the samples that had not experienced a historical PAH contamination, these pollutants could cause damages to the structure and composition of the microbial community, which could decrease the evenness of it. From the results of beta diversity analysis, when the studied variable is set as “PAH concentration” or “historical PAH contamination”, it could be found that there are some differences on the structure and composition of microbial communities in the samples between different groups. This could also result from the effects on microorganisms caused by PAHs, as there are significant differences on the tolerance and utilization of PAHs between different microorganisms, which finally led to the variations on the structure and composition of the microbial communities. Especially when the studied variable is “historical PAH contamination”, it could be observed that the microbial communities in the samples between two groups were quite different, which could further reflect the impact on the microbial diversity in the coal-tar contaminated media caused by this factor. Moreover, according to the figures with the variable of “environmental material”, it could be obtained that there is a similarity on the structure and composition of the microbial communities in most of samples, even for the samples belong to different groups, which could indicate that the effects caused by these contaminants on the microbial communities in different media might be similar. From the results of environmental filtering analysis, when the studied variable is “PAH concentration”, a trend could be observed, which is with the increase of PAH concentration in samples, the value of NRI/NTI became larger. This could indicate that there is a strong environmental pressure imposed by PAHs in the samples, and further proved the biological toxicity of PAHs and the effects on the microbial communities

caused by them. Besides, it could also be found that for some samples with low PAH concentration, the values of NRI/NTI are even less than 0, which could show a relatively strong effect of competitive exclusion compared with the environmental filtering. This finding could also indicate that when the PAH concentration in the environment is low (does not cause damages to the microbial community in the environment), the PAHs might play a role as the microbial carbon source and then intensify the competition between different microbial populations in some ways, which could also demonstrate the bio-accessibility of PAHs that mentioned in the previous studies. However, when the studied variable switched to historical PAH contamination, the result of environmental filtering analysis, which the values of NRI/NTI of the samples that had experienced the historical PAH contamination are higher than the samples had not, indicating that compared with the samples without historical contamination, the samples with it might bear a relatively greater environmental pressure, which could be considered as inconsistent with the expected result. This may result from the differences on some other factors between samples, such as natural environmental conditions, or the evolution stage of the microbial communities in samples. Considering the large gap between the number of samples contained in two groups, this could also be considered as a possible reason for this result.

From the results of taxa bars analysis, it is obvious that the compositions of microbial communities between the samples extracted from soil and seawater are similar in a certain extent, while significant differences could be observed on the specific proportions occupied by each species. Except for the huge differences on the features between two types of media, this might also result from the differences on some other factors such as temperature, salinity or nutrient composition between two groups of samples. Additionally, when the studied variable is set as "PAH concentration", it could be shown that with a relatively high concentration of PAHs, both the richness and evenness of the microbial community would be declined, which could also indicate a relatively strong environmental pressure caused by PAHs with high concentration. Moreover, obvious differences on both the structure and composition of the microbial community between the samples that had experienced historical PAH contamination and the samples had not could also be observed, which means the previous PAH contamination might have eliminated parts of microorganisms that living in the original environment, and this could further prove the biological toxicity of PAHs on the microorganisms. In the meantime, several species, which have been regarded as the bio-degrader of PAHs, such as *Bacillus*, *Pseudomonas* and *Sphingobium*, were found in almost all the samples contained in this project, some of them even occupied a considerable proportion in the microbial communities. This could further demonstrate the possibility of utilizing the microorganisms in the environment to degrade the PAH contaminants, which has been mentioned many times by previous studies.

5 Conclusion

In summary, several points could be obtained from the analysis above. To some extent, the PAHs left in the media, which were contaminated by coal tar, could lead to a two-sides effect on the microorganisms that living in the environment. On the one hand, it could be considered that the PAH has a certain biological toxicity to the microbial community in the environment, high concentration of PAHs would significantly reduce both the richness and evenness of the microbial community, and impose a relatively strong environmental pressure on the microorganisms. However, on the other hand, PAHs could also be used as a carbon source for the microorganisms, which could provide the possibility for the biodegradation of PAHs in some ways, but this process might be limited by a series of factors such as the specific types and concentration of PAHs. In the meantime, it could also be found that whether the samples had experienced PAH contamination before has a significant impact on the microbial diversity in them, which is consistent with the results of previous studies. Moreover, some species of microorganisms, which could be utilized in the process of biodegradation of PAHs and the bioremediation of contaminated media were found in the samples according to related results. This could further indicate the utilization of the biodegradation on PAHs and some other pollutants has a relatively general applicability.

However, there are still some limitations in the analysis process of this project. Based on some detailed information about the samples contained in the analysis, it could be inferred that the isolation source of the samples might also result in some influences on the microbial diversity in the environment, more detailed studies on this point might be required. Besides, in the analysis process of this project, for various studied variables, the noticeable differences on the number of samples that contained in different groups may also interfere the results of analysis to a certain extent, which should be avoided in similar studies. In addition, the setting of the variable of pH value is not successful, since the samples with $\text{pH} > 7$ are all extracted from seawater, and the samples with $\text{pH} < 7$ are mostly soil samples, which leads to the results on the analysis focused on the “environmental material” and “pH value” are highly overlapping. This also means that both of these two variables did not fully exert their effects during the analysis.

References

- Abdel-Shafy, H. and Mansour, M., 2016. A review on polycyclic aromatic hydrocarbons: Source, environmental impact, effect on human health and remediation. *Egyptian Journal of Petroleum*, 25(1), pp.107-123.
- Bae, H., Huang, L., White, J., Wang, J., DeLaune, R. and Ogram, A., 2018. Response of microbial populations regulating nutrient biogeochemical cycles to oiling of coastal saltmarshes from the Deepwater Horizon oil spill. *Environmental Pollution*, 241, pp.136-147.
- Bell, T., Stefani, F., Abram, K., Champagne, J., Yergeau, E., Hijri, M. and St-Arnaud, M., 2016. A Diverse Soil Microbiome Degrades More Crude Oil than Specialized Bacterial Assemblages Obtained in Culture. *Applied and Environmental Microbiology*, 82(18), pp.5530-5541.
- Benjjneb.github.io. 2020. Index.Utf8.Md. [online] Available at: <<https://benjjneb.github.io/dada2/index.html>> [Accessed 21 August 2020].
- Bolyen E, Rideout JR, Dillon MR, Bokulich NA, Abnet CC, Al-Ghalith GA, Alexander H, Alm EJ, Arumugam M, Asnicar F, Bai Y, Bisanz JE, Bittinger K, Brejnrod A, Brislawn CJ, Brown CT, Callahan BJ, Caraballo-Rodríguez AM, Chase J, Cope EK, Da Silva R, Diener C, Dorrestein PC, Douglas GM, Durall DM, Duvallet C, Edwardson CF, Ernst M, Estaki M, Fouquier J, Gauglitz JM, Gibbons SM, Gibson DL, Gonzalez A, Gorlick K, Guo J, Hillmann B, Holmes S, Holste H, Huttenhower C, Huttley GA, Janssen S, Jarmusch AK, Jiang L, Kaehler BD, Kang KB, Keefe CR, Keim P, Kelley ST, Knights D, Koester I, Kosciulek T, Kreps J, Langille MGI, Lee J, Ley R, Liu YX, Loftfield E, Lozupone C, Maher M, Marotz C, Martin BD, McDonald D, McIver LJ, Melnik AV, Metcalf JL, Morgan SC, Morton JT, Naimey AT, Navas-Molina JA, Nothias LF, Orchanian SB, Pearson T, Peoples SL, Petras D, Preuss ML, Priesse E, Rasmussen LB, Rivers A, Robeson MS, Rosenthal P, Segata N, Shaffer M, Shiffer A, Sinha R, Song SJ, Spear JR, Swafford AD, Thompson LR, Torres PJ, Trinh P, Tripathi A, Turnbaugh PJ, Ul-Hasan S, van der Hooft JJJ, Vargas F, Vázquez-Baeza Y, Vogtmann E, von Hippel M, Walters W, Wan Y, Wang M, Warren J, Weber KC, Williamson CHD, Willis AD, Xu ZZ, Zaneveld JR, Zhang Y, Zhu Q, Knight R, and Caporaso JG. 2019. Reproducible, interactive, scalable and extensible microbiome data science using QIIME 2. *Nature Biotechnology* 37: 852–857. <https://doi.org/10.1038/s41587-019-0209-9>
- Brown, B., LePrell, R., Franklin, R., Rivera, M., Cabral, F., Eaves, H., Gardiakos, V., Keegan, K. and King, T., 2015. Metagenomic analysis of planktonic microbial consortia from a non-tidal urban-impacted segment of James River. *Standards in Genomic Sciences*, 10(1).
- Brown, D., Gupta, L., Kim, T., Keith Moo-Young, H. and Coleman, A., 2006.

Comparative assessment of coal tars obtained from 10 former manufactured gas plant sites in the Eastern United States. *Chemosphere*, 65(9), pp.1562-1569.

Canet, R., Birnstingl, J., Malcolm, D., Lopez-Real, J. and Beck, A., 2001. Biodegradation of polycyclic aromatic hydrocarbons (PAHs) by native microflora and combinations of white-rot fungi in a coal-tar contaminated soil. *Bioresource Technology*, 76.

Collins, M., Williams, P. and McIntosh, D., 2001. Ambient Air Quality at the Site of a Former Manufactured Gas Plant. *Environmental Monitoring and Assessment*, 68(2), pp.137-152.

Cran.r-project.org. 2020. Analyses Of Phylogenetics And Evolution [R Package Ape Version 5.4-1]. [online] Available at: <<https://cran.r-project.org/web/packages/ape/>> [Accessed 21 August 2020].

Cran.r-project.org. 2020. CRAN - Package Vegan. [online] Available at: <<https://cran.r-project.org/web/packages/vegan/>> [Accessed 21 August 2020].

Cran.r-project.org. 2020. Create Elegant Data Visualisations Using The Grammar Of Graphics [R Package Ggplot2 Version 3.3.2]. [online] Available at: <<https://cran.r-project.org/web/packages/ggplot2/>> [Accessed 21 August 2020].

Cran.r-project.org. 2020. Extension Of 'Data.Frame' [R Package Data.Table Version 1.13.0]. [online] Available at: <<https://cran.r-project.org/web/packages/data.table/>> [Accessed 21 August 2020].

DeBruyn, J., Chewning, C. and Sayler, G., 2007. Comparative Quantitative Prevalence of Mycobacteria and Functionally Abundant *nidA*, *nahAc*, and *nagAc* Dioxygenase Genes in Coal Tar Contaminated Sediments. *Environmental Science & Technology*, 41(15), pp.5426-5432.

Dionisi, H., Chewning, C., Morgan, K., Menn, F., Easter, J. and Sayler, G., 2004. Abundance of Dioxygenase Genes Similar to *Ralstonia* sp. Strain U2 *nagAc* Is Correlated with Naphthalene Concentrations in Coal Tar-Contaminated Freshwater Sediments. *Applied and Environmental Microbiology*, 70(7), pp.3988-3995.

Docs.qiime2.org. 2020. What Is QIIME 2? — QIIME 2 2020.6.0 Documentation. [online] Available at: <<https://docs.qiime2.org/2020.6/about/>> [Accessed 21 August 2020].

Environmental Science & Technology, 1994. REMEDIATING TAR-CONTAMINATED SOILS AT MANUFACTURED GAS PLANT SITES. 28(6), pp.266A-276A.

Enzminger, J. and Ahlert, R., 1987. Environmental fate of polynuclear aromatic hydrocarbons in coal tar. *Environmental Technology Letters*, 8(1-12), pp.269-278.

Franck, H., 1963. THE CHALLENGE IN COAL TAR CHEMICALS. *Industrial & Engineering Chemistry*, 55(5), pp.38-44.

Gauchotte-Lindsay, C., Aspray, T., Knapp, M. and Ijaz, U., 2019. Systems biology approach to elucidation of contaminant biodegradation in complex samples – integration of high-resolution analytical and molecular tools. *Faraday Discussions*, 218, pp.481-504.

Hatheway, A., 2012. Remediation Of Former Manufactured Gas Plants And Other Coaltar Sites. Boca Raton, FL: CRC Press.

Humans, I., 2012. Chemical Agents And Related Occupations. [Place of publication not identified]: International Agency for Research on Cancer.

Joey711.github.io. 2020. Phyloseq: Explore Microbiome Profiles Using R. [online] Available at: <<https://joey711.github.io/phyloseq/>> [Accessed 21 August 2020].

Johnsen, A. and Karlson, U., 2007. Diffuse PAH contamination of surface soils: environmental occurrence, bioavailability, and microbial degradation. *Applied Microbiology and Biotechnology*, 76(3), pp.533-543.

Júlio, A., de Cássia Mourão Silva, U., Medeiros, J., Morais, D. and dos Santos, V., 2019. Metataxonomic analyses reveal differences in aquifer bacterial community as a function of creosote contamination and its potential for contaminant remediation. *Scientific Reports*, 9(1).

Jung, M., Kim, J., Sinninghe Damsté, J., Rijpstra, W., Madsen, E., Kim, S., Hong, H., Si, O., Kerou, M., Schleper, C. and Rhee, S., 2016. A hydrophobic ammonia-oxidizing archaeon of the Nitrosocosmicus clade isolated from coal tar-contaminated sediment. *Environmental Microbiology Reports*, 8(6), pp.983-992.

Lee, P., Ong, S., Golchin, J. and Nelson, G., 2001. Use of solvents to enhance PAH biodegradation of coal tar-. *Water Research*, 35(16), pp.3941-3949.

Li, J., Pignatello, J., Smets, B., Grasso, D. and Monserrate, E., 2005. BENCH-SCALE EVALUATION OF IN SITU BIOREMEDIATION STRATEGIES FOR SOIL AT A FORMER MANUFACTURED GAS PLANT SITE. *Environmental Toxicology and Chemistry*, 24(3), p.741.

McGregor, L., 2012. Environmental Forensic Investigation of Coal Tars from Former Manufactured Gas Plants.

McGregor, L., Gauchotte-Lindsay, C., Nic Daéid, N., Thomas, R. and Kalin, R., 2012. Multivariate Statistical Methods for the Environmental Forensic Classification of Coal Tars from Former Manufactured Gas Plants. *Environmental Science & Technology*, 46(7), pp.3744-3752.

Nearing, J., Douglas, G., Comeau, A. and Langille, M., 2018. Denoising the Denoisers: an independent evaluation of microbiome sequence error-correction approaches. *PeerJ*, 6, p.e5364.

Neethu, C., Saravanakumar, C., Purvaja, R., Robin, R. and Ramesh, R., 2019. Oil-Spill Triggered Shift in Indigenous Microbial Structure and Functional Dynamics in Different Marine Environmental Matrices. *Scientific Reports*, 9(1).

Prodan, A., Tremaroli, V., Brolin, H., Zwinderman, A., Nieuwdorp, M. and Levin, E., 2020. Comparing bioinformatic pipelines for microbial 16S rRNA amplicon sequencing. *PLOS ONE*, 15(1), p.e0227434.

Ren, G., Ren, W., Teng, Y. and Li, Z., 2015. Evident bacterial community changes but only slight degradation when polluted with pyrene in a red soil. *Frontiers in Microbiology*, 6.

Rstudio.com. 2020. Rstudio. [online] Available at: <<https://rstudio.com/products/rstudio/>> [Accessed 21 August 2020].

Sperfeld, M., Rauschenbach, C., Diekert, G. and Studenik, S., 2018. Microbial community of a gasworks aquifer and identification of nitrate-reducing *Azoarcus* and *Georgfuchsia* as key players in BTEX degradation. *Water Research*, 132, pp.146-157.

Tatariw, C., Flournoy, N., Kleinhuizen, A., Tollette, D., Overton, E., Sobecky, P. and Mortazavi, B., 2018. Salt marsh denitrification is impacted by oiling intensity six years after the Deepwater Horizon oil spill. *Environmental Pollution*, 243, pp.1606-1614.

Thavamani, P., Megharaj, M. and Naidu, R., 2011. Multivariate analysis of mixed contaminants (PAHs and heavy metals) at manufactured gas plant site soils. *Environmental Monitoring and Assessment*, 184(6), pp.3875-3885.

Valencia-Agami, S., Cerqueda-García, D., Putzeys, S., Uribe-Flores, M., García-Cruz, N., Pech, D., Herrera-Silveira, J., Aguirre-Macedo, M. and García-Maldonado, J., 2019. Changes in the Bacterioplankton Community Structure from Southern Gulf of Mexico During a Simulated Crude Oil Spill at Mesocosm Scale. *Microorganisms*, 7(10), p.441.

Wexler, P., 2014. *Encyclopedia Of Toxicology*. Amsterdam: Elsevier.

Williams, H., Bigby, M., Herxheimer, A., Naldi, L., Rzany, B., Dellavalle, R., Ran, Y. and Furue, M., 2014. Evidence-Based Dermatology. Hoboken: Wiley.

Wu, M., Guo, X., Wu, J. and Chen, K., 2020. Effect of compost amendment and bioaugmentation on PAH degradation and microbial community shifting in petroleum-contaminated soil. *Chemosphere*, 256, p.126998.