Coursework Declaration and Feedback Form

The Student should complete and sign this part

Student Number: 2807941L	Student Name: Yi Liu
Programme of Study (e.g. MSc in Electronics and Electrical Engineering): MSc in Computer systems Engineering	
Course Code: ENG5059P	Course Name: MSc Project
Name of **First** Supervisor: Dr Umer Zeeshan Ijaz	Name of **Second** Supervisor:
Title of Project: Exploring microbial interactions and diversity in contaminated soil	

Declaration of Originality and Submission Information

I affirm that this submission is all my own work in accordance with the University of Glasgow Regulations and the School of Engineering requirements Signed (Student) : Yi Liu	
Date of Submission : 14 August 2023	

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

Exploring microbial interactions and diversity in contaminated soil

Yi Liu

2807941L

Supervised by Dr. Umer Zeeshan Ijaz

August 14, 2023

COMPUTER SYSTEMS ENGINEERING MSc

Abstract

Soil pollution can be understood as the soil containing more harmful components than the normal level, thus affecting the normal ecology and natural function of the soil. This problem is still an urgent challenge to global environmental protection and sustainable development. In addition to affecting soil ecology, the spread of soil pollution elements may lead to more serious environmental problems due to the spread of toxins to the air or nearby waters. In recent years, the potential of biodegradation has made it new technical support to solve soil pollution. Among the many soil pollutants, polycyclic aromatic hydrocarbons (PAHs) are one of the lipophilic organic pollutants commonly in areas affected by modern industrial activities such as industrial chemical industry, urban traffic and garbage incineration. Because of the universality and harmfulness of PAHs, it is particularly important to monitor and study PAHs and other organic pollutants to find ways to solve the problem of soil pollution. In summary, this experiment chose PAHs as the research goal of soil pollutants to study the temporal dynamics of PAHs and microbial communities in the soil environment. By tracking the changes in microbial communities, we can better understand their interaction in the process of biodegradation and the mechanism of microbial communities so as to provide key information for soil pollution response and remediation process and optimize soil remediation strategies in the future.

In this experiment, we will comprehensively use high-resolution analysis and molecular tool biology methods to process and analyze the biological information of soil samples of COV and CH collected from the United Kingdom and Switzerland. According to the composition information obtained from the two samples, both samples were detected to have different degrees of polycyclic aromatic hydrocarbons (PAHs) pollution.

The RStudio software will analyze the 16SrRNA sequencing data and chemical and molecular metadata obtained from the sample points in this experiment. Draw lessons from the research progress of biodegradation methods of Gauchotte-Lindsay 's team (2019), and on this basis, formulate and discover new research strategies to study the time dynamics of microbial communities, reveal the interaction between environmental factors and biological communities, and the symbiotic relationship between microorganisms. This study will use zeta diversity and network analysis as the main methods to provide valuable insights into the role and mechanism of microorganisms in the process of biodegradation. This new study will contribute to a better understanding of the biodegradation processes involving microorganisms and provide the basis for developing more effective soil pollution control strategies.

Keywords: Soil pollution, Biodegradation, Polycyclic aromatic hydrocarbons, Zeta diversity, Network, Microbial community

Acknowledgement

In the process of completing this paper, I sincerely thank everyone for their help and support during this process.

Firstly, I would like to express my deep respect and gratitude to my mentor Professor Umer Zeeshan Ijaz. His courses and guidance every Thursday are full of patience and professionalism, and he has invested a lot of time answering our questions and caring about the progress of our project research. Professor Ijaz's guidance and persistent pursuit of knowledge have greatly benefited me.

At the same time, I would like to thank the other volunteer teaching assistants in the laboratory. Especially Ms. Kelly. J.Stewart's selfless help helped me overcome many difficulties in my paper writing. Ms. Aqsa.Ameer provided valuable suggestions for my academic poster, further improving my academic expression. I would also like to express my special gratitude to Ms. Uzma for her selfless help in my learning of R language programming. Even though I have difficulties expressing myself in the language, she has always been patient and has played a great role in promoting my learning. In addition, I would like to thank the other students in our laboratory for learning together, exploring together, and overcoming difficulties together. Their help and support have made me more determined on my academic path.

Finally, I would like to express my deep gratitude to my family and friends. Their companionship and support have made me feel immensely warm and loving in my study abroad life in the UK and have also given me the strength to move forward on my academic path.

Thank you again to all the people who have helped me on my way to study. You have filled me with confidence and motivation, enabling me to complete this research.

List of Abbreviations

PAHs	Polycyclic Aromatic Hydrocarbons
COV	Sampling points for contaminated soil in UK
CH	Sampling points for contaminated soil in Switzerland
RNA	Ribonucleic Acid
DNA	Deoxyribonucleic Acid
POPs	Persistent Organic Pollutants
PCB	Polychlorinated Biphenyls
PCDD	Polychlorinated Dibenzodioxins
OCP	Organochlorine Pesticides
IARC	International Agency for Research on Cancer
PCR	Polymerase Chain Reaction
qPCR	Quantitative Polymerase Chain Reaction
OTU	Operational Taxonomic Unit
LOI	Loss on Ignition
UCLUST	Software for clustering biological sequences
QIIME	Quantitative Insights Into Microbial Ecology
IVI	Integrated Value of Influence

List of Figures

Contents

1 Introduction

1.1 Research Background

Global environmental issues have become the focus of global public attention, with soil pollution being particularly prominent, which directly affects human health and the stability of ecosystems (Mirsal, 2008). There is a wide range of pollution sources, including but not limited to industrial waste, excessive use of pesticides and fertilizers, leakage of petrochemical products, untreated urban and rural garbage, and mining activities (Mishra et al., 2016). These activities accumulate heavy metals, organic matter, radioactive substances, and other harmful compounds in the soil, which may lead to a decrease in soil fertility, affect crop growth, and potentially enter the human body through the food chain, posing a threat to human health (Mirsal&Mirsal, 2008).

Although traditional soil remediation methods (including physical, chemical, and thermal methods) have some effectiveness, there are many problems (Sun et al., 2018). For example, physical and chemical methods such as pyrolysis and incineration may generate secondary pollution and pose environmental risks (Ahmad et al., 2020). Traditional remediation methods, such as soil cover, are expensive and have limited effectiveness in the remediation of large-scale contaminated sites (Liu et al., 2018). Chemical restoration methods such as chemical reduction and oxidation require a large amount of chemical reagents. They are difficult to avoid the generation of side reactions and other harmful products during the repair process (Lim et al., 2016).

Therefore, more environmentally friendly and economical remediation methods have become the focus of research, with bioremediation technology demonstrating enormous potential and advantages as a sustainable and low-cost soil remediation method (Ayangbenro&Babalola, 2017). By using biological indicators to evaluate soil quality, interpret microbial population density, or measure basic biological activities, reference can be provided for the further development of bioremediation technology (Margesin&Schinner, 2005).

Currently, many studies have focused on the transformation of typical persistent organic pollutants (POPs) in soil, such as polychlorinated biphenyls (PCBs), polychlorinated dibenzo p-dioxins (PCDDs), and organochlorine pesticides (OCPs) (Abhilash et al., 2013). However, research on polycyclic aromatic hydrocarbons (PAHs) is relatively limited. PAHs are highly persistent, toxic, and widely present environmental pollutants, which are by-products of incomplete combustion of various organic compounds. They can be used as particulate-related air pollutants or directly emitted as volatile pollutants from various human and natural sources (Kuppusamy et al., 2017). Once deposited into the soil from the air, PAHs accumulate for a longer period of time, with half-lives of 83-193 days for phenanthrene and 282-535 days for benzopyrene, respectively (Wang et al., 2008).

Although biodegradation shows its potential for soil remediation, the current biodegradation technology is still not perfect as to how to ensure the stability and sustainability of the key strains of soil remediation in the real soil environment and how to avoid the potential risks of key species strains to the local environment and ecology. Therefore, for persistent organic pollutants such as PAHs, the research should focus on their migration and transformation mechanism in the environment. At the same time, the study on the interaction and diversity of microorganisms in the bioremediation process will also provide a basis for us to formulate more targeted and safe remediation strategies.

1.2 Lipophilic organic pollutants polycyclic aromatic hydrocarbons

Polycyclic aromatic hydrocarbons (PAHs) are a special kind of organic compound. It is unique in that its structure is composed of two or more aromatic rings closely connected. Typical polycyclic aromatic hydrocarbons include naphthalene, dibenzo anthrone and so on. In the natural environment, organic compounds such as PAHs can be broken down and transformed by many microorganisms, especially if appropriate environmental conditions are present and sufficient nutrients are available. These microorganisms can use PAHs as a source of carbon and energy for biodegradation. However, it should be pointed out that not all PAHs are equally easy to be decomposed by microorganisms. The aromatic ring is a highly conjugated structure, and its special electron mobility ensures the stability of its chemical structure. PAHs with simple structures and low molecular weight (such as naphthalene) are relatively easy to be decomposed. However, PAHs with high molecular weight and complex structure (such as benzo [a] pyrene) are more difficult to biodegrade.

Natural sources and anthropogenic sources are the main sources generally recognized in the current research on PAHs.(Mojiri, Zhou, Ohashi, Ozaki, & Kindaichi, 2019). Natural sources are mainly due to natural phenomena such as forest fires, volcanic eruptions and crustal movements (Abdel-Shafy&Mansour,2016). Anthropogenic sources are mainly from anthropogenic activities such as coal burning, petroleum, wood burning, and traffic exhaust (Ravindra et al.,2008). At the same time, in modern industrial production processes caused by waste or metallurgical products, all kinds of organic matter in the process of incomplete combustion often produce PAHs (Choi et al., 2010). Human activity is a major driver of increasing PAHs concentrations in the environment (Patel et al.,2020).

The solubility of PAHs in oily or non-polar environments is much higher than that in aqueous or polar environments, which depends on their stable aromatic hydrocarbon composition (Gan et al.,2009). This property makes the distribution and migration

pattern of PAHs in the environment special. Therefore, aerosol particles in the air are the main means of atmospheric migration of PAHs, while in soil and water, their presence changes to adsorption on the surface of organic matter and particles (Samanta et al.,2002).

In addition, PAHs exist in a variety of forms, including gaseous and granular forms. Most PAHs exist mainly as particles due to their low volatility. However, some low-molecular-weight PAHs (such as naphthalene and phenylene) can exist as a gas because of their small molecular weight. In contrast, high molecular weight PAHs (such as benzopyrene) exist mainly in granular form due to their large molecular weight and low volatility.PAHs in granular form can enter soil and water through atmospheric deposition and persist in them for a long time (Freman &Cattell,1990). Moreover, in the process of biodegradation of PAHs, some intermediate metabolites may be generated. These intermediates may, in some cases, be more toxic or more difficult to break down than the original PAHs.And these toxic intermediate metabolites, as well as polycyclic aromatic organic compounds that are not biodegradable, often inhibit microbial development and reduce their ability to break down pollutants. When microorganisms cannot completely decompose pollutants, the soil will form permanent pollution, causing great harm to the environment and organisms.

It should be noted that PAHs have significant biological toxicity and persistence. They not only have toxicity to aquatic and soil organisms but also undergo bioconcentration and biomagnification in the food chain. Among them, some PAHs (such as naphthalene, phenanthrene, or benzene) have been confirmed as human carcinogens by the International Agency for Research on Cancer (IARC) (Ghost, Ghost, Dutta,&Ahn, 2016).

However, although the environmental impact and toxicity issues of PAHs have received widespread attention, there are still many challenges in research on their

behaviour in the environment, biodegradation mechanisms, and health risk assessment. Therefore, in-depth research and understanding of PAHs is crucial, which will help us better understand and manage PAHs pollution issues.

1.3 Microbial gene sequencing methods

Genome sequencing technology plays a crucial role in environmental and Microbio -logical research. In the early days, traditional gene sequencing methods mainly relied on clone library construction and Sanger sequencing technology, which significantly contributed to studying the genomes of a single species in the environment. However, their high cost, time consumption, and low coverage of complex microbial communities limited their application in environmental microbial community analysis (Fran ç a, Carrilho,&Kist, 2002). However, with the emergence of second-generation high-throughput sequencing technologies, including Illumina sequencing and Ion Torrent sequencing, they provide higher sequencing depth and wider coverage, providing new possibilities for the study of environmental microbial communities (Park&Kim, 2016).

Among these methods, 16S rRNA gene sequencing is the most commonly used method for microbial classification and identification. The 16S rRNA gene is a conserved gene in bacteria and archaea, and the sequence of its mutated region can be used to distinguish different microbial species (Janda&Abbott, 2007).By conducting high-throughput sequencing of the 16S rRNA gene, we can obtain detailed information on the composition of microbial communities in environmental samples. Specifically, as shown in Figure 1, by designing specific primers for PCR amplification of the 16S rRNA gene, the obtained amplicons were sequenced, and then bioinformatics analysis was performed to identify the types and relative abundance of microorganisms present in the sample (Abellan Schneiyder et al., 2021).

The 16S rRNA sequencing method has extensive applications in environmental and microbiology. 16S rRNA gene sequencing is commonly used for the identification, classification, and quantification of microorganisms in complex biological mixtures, as shown in Figure 2 (Bowman&Kwon, 2016). It can be used to study the composition and dynamic changes of soil, water, air, and human microbial

communities, as well as the impact of environmental factors such as pollution, temperature, humidity, etc., on microbial communities. The advantage of this method lies in its ability to accurately, quickly, and economically analyze the composition of microbial communities in complex environmental samples.

We found significant differences in the composition and abundance of microbial communities among soil samples with different levels of pollution. This difference may be related to the degradation ability and tolerance of microorganisms to pollutants. Therefore, 16S rRNA gene sequencing will provide us with the main data information for understanding the structure and function of microbial communities in polluted soil, as well as the role of microorganisms in pollutant degradation.

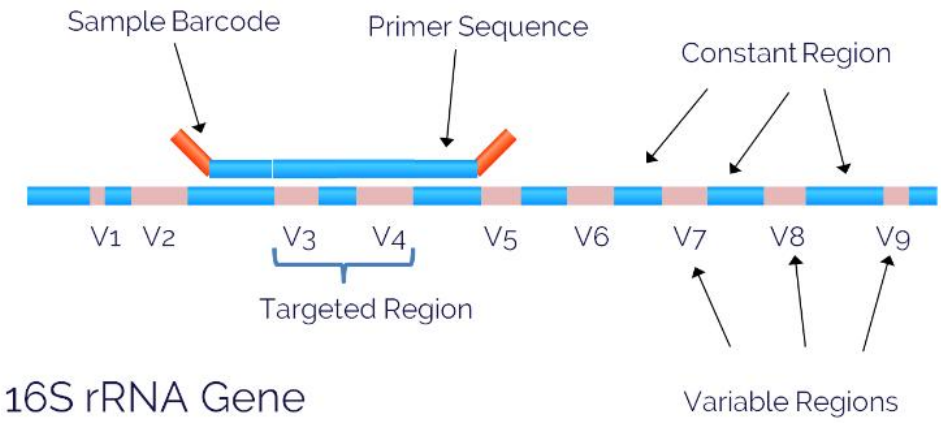


Figure 1: 16s rRNA gene composition (LC Sciences,2023)

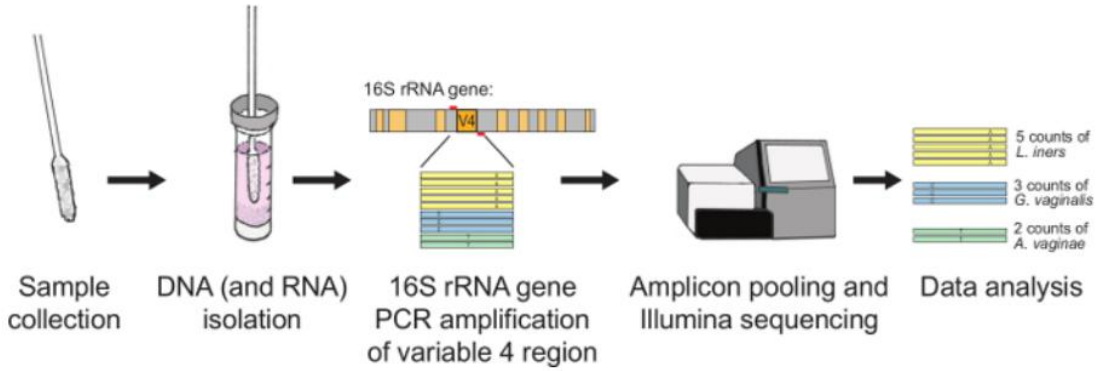


Figure 2: 16S rRNA gene sequencing for human bacterial community identification (Bowman, B.A., 2016)

1.4 Data acquisition methods for microbial communities

In order to reveal detailed information on microbial communities in contaminated soil samples of CH and COV, this study employed the Operational Taxonomy Unit (OTU) technique. OTU is a population analysis method widely used in the fields of ecology and microbiology, and its accuracy has significant advantages in showcasing the composition and diversity of microbial communities (Llad ó Fern á ndez, V ě Trovsk ý, & Baldrian, 2019).

The working principle of OTU is based on the similarity of 16S rRNA gene sequences. Usually, if two or more 16S rRNA gene sequences exhibit similarity at a level of 97% or higher, they are classified as the same OTU (Jackson et al., 2016). This method not only has the ability to identify species but also has relatively low computational and storage costs when processing a large number of samples. Therefore, OTU technology has wide applications in fields such as environmental microbiology, human microbiome research, and soil microbiology (Nguyen et al., 2016).

In this study, OTU technology was used for high-throughput sequencing of microbial DNA in CH and COV samples to quickly identify microbial species in soil samples and obtain their relative abundance. By establishing OTU, the microbial communities in different soil samples can be compared on the same platform, which is convenient for diversity analysis and microbial community structure comparison.

Another significant advantage of OTU is that it can quantitatively compare the differences of microbial communities under different samples or treatment conditions. This advantage makes it possible to track the temporal dynamics of microbial communities in contaminated soil samples, and by comparing the differences between microbial communities in different samples, we can further reveal the interaction of microbial communities under different soil conditions (Preheim et al.,2013).

1.5 Purpose and Objectives

This study aimed to investigate the composition of microbial communities in contaminated soil samples collected from COV in the United Kingdom and CH in Switzerland by analyzing 16SrRNA sequencing data and its accompanying metadata. Data analysis reveals the temporal dynamics of microbial communities under environmental impact and the interaction between contaminated soil and microbial communities. It provides a new idea for using biological explanation methods to solve the soil pollution problem.

2 Methodology

2.1 Sample Description

The project uses soil samples collected by Professor UmerZ.Ijaz's team and their collaborators from two different original natural gas plants. This study, 26 soil samples were collected, of which 17 were collected from COV sites in the United Kingdom, and nine were collected from CH sites in Switzerland (Gauchotte-Lindsay et al.,2019).

The soil samples are stored in plastic tubes with a temperature of 4 degrees Celsius to reduce interference from external conditions. In the records of this study, 17 soil samples collected from COV sites in the UK typically presented dark colours, dense textures, and were slightly sticky. Unlike the soil samples from COV sites, the soil samples from CH sites were brown and relatively dry. According to the different physical states of the two samples, different filtering methods were selected. 17 samples from the COV site were filtered once through a 10mm sieve, while 9 samples from the CH site were filtered through 1.7, 2.36, and 10mm sieves, respectively.

For each sample collected in this study, the moisture content and loss on ignition (LOI) of each sample were monitored by storing the sub-samples in incubators at different temperatures (Gauchotte Lindsay et al., 2019). Based on the data collected from the above soil samples, genomic DNA was extracted twice from two sites, one for a specific DNA sequence using qPCR to quickly detect specific microbial communities such as alkB12 in the soil and evaluate their biodegradation potential for hydrocarbon compounds. Afterwards, 16S rRNA sequencing was used again to provide more comprehensive information on microbial communities, including their diversity and composition. Fully utilize their respective advantages to better evaluate the biodegradation potential in soil.

2.2 Biological Gene Data Extraction

After processing multiple soil samples from different regions, we need to extract biological gene data from the microorganisms in the soil samples. This study chose 16S rRNA gene sequencing to collect gene data from the samples. This technology is particularly applicable in bacterial gene sequencing, as 16S rRNA is difficult to change during the long evolutionary process of bacteria, making it a reliable method for measuring bacterial evolution. Meanwhile, the gene sequence length can be used for informatics analysis. The characteristic of this technology is that it achieves differentiation of different microbial species by amplifying and proliferating fragments of the V3 and V4 regions of the 16S rRNA gene (Janda&Abbott, 2007). The entire operation process includes steps such as DNA extraction, PCR amplification, sequencing, and subsequent data analysis.

Firstly, extract microbial DNA from soil samples. Amplify the 16S rRNA gene using PCR technology to obtain sufficient DNA material for sequencing. During the amplification process, 16S rRNA universal primers capable of covering most microorganisms were used (Yoon et al., 2017). After sequencing on the sequencing platform, a large amount of sequencing data was obtained. These data were originally formed into countless DNA fragments, each of which originated from the 16S rRNA gene of a certain microorganism. Afterwards, these fragments were compared with known 16S rRNA databases using bioinformatics software and classified into different operational taxonomic units (OTUs) based on similarity. OTU is a non-human sequence cluster constructed based on sequence similarity using UCLUST software and Quantitative Insights Into Microbial Ecology (QIIME) software specifically designed for high-throughput 16S rRNA sequencing research (Statnikov et al., 2013) before processing raw DNA sequencing data. Each OTU represents a possible microbial species.

Finally, an OTU table was generated based on the quantity and abundance of different

OTUs in each sample. OTU table provides a comprehensive and detailed description of microbial species and relative abundance and an essential basis for subsequent biodiversity and functional analysis. In this study, Rstudio was used to analyze the data and parameters of OTU species to determine the dominant members who played an essential role in the different microbial communities in the soil samples of the two sites.

2.3 Statistical Analysis

In this study, statistical data analysis was mainly completed using R Studio. We have organized and filtered the required tables and data according to the previously set preparation method. These data include various variables and parameters required for the study. Using R studio for analysis, we performed key tasks, including descriptive statistical analysis, inferential statistical analysis, and data visualization.

In addition, we also analyzed the metadata related to the research. The metadata analysis helps us have a deeper understanding of the background and structure of the data, thereby ensuring the accuracy and reliability of the analysis.

2.3.1 zeta diversity

In the study of microbial diversity in ecosystems, zeta diversity method is a new multi-sample and multi-time dynamic research method, which is used to describe the shared number of species among multiple communities. This parameter can be used to study the transition of microorganisms and distinguish diversity patterns in rare groups, and infer the deterministic and random drivers of the community structure. This parameter can be used to predict and analyze biodiversity patterns and their responses to environmental changes. Therefore, zeta diversity can be understood as a unified concept and measurement method (Hui&McGeoch,2014) based on diversity measurement, patterns and relationships of species occurrence.

Zeta order, that is, the number of samples compared together in the process of species diversity research. With the increase of the number of contrasted samples, the average number of shared groups among samples decreases, while the contribution of more and more common groups to zeta diversity increases. The rate and trend changes of zeta diversity provide information on community structure and inference of the process of promoting community aggregation. Therefore, the analysis of zeta diversity is often accompanied by mathematical regression methods such as exponential regression and power law regression. In this study, we used the degree of fitting between the two regression analysis methods and the dynamic curve of zeta diversity to infer the time dynamics and community structure of microbial communities in soil samples.

When the change of zeta diversity is more in line with the exponential form, it shows that with the increase of zeta order, the probability of common and rare OTU is similar, and the main reason for species replacement is random or diffusion limitation of the community itself. When the dynamic curve is more similar to the power law form, it shows that with the increase of sample points, the probability of detecting OTU of common species is higher than that of rare species, and the main reason for

this community change may be the deterministic process caused by soil, climate and other environment.

Compared with previous diversity analysis methods and zeta diversity methods, in microbiology, Alpha diversity analyzes species richness (Thukral,2017) by analyzing individual samples or biological samples in a local homogeneous environment, while Beta diversity focuses on measuring species composition differences between different samples or communities (Legendre,2008). Zeta diversity is on this basis to do in-depth. The main exploration is the co-occurrence model among species (Hui&McGeoch,2014).

Zeta diversity stems from the need for a finer and deeper understanding of biodiversity, and its core concept is to consider the co-occurrence model among species. Therefore, Zeta diversity has a unique sensitivity to the interaction and organizational structure of biological communities. This more in-depth exploration provides a new quantifiable method for soil remediation. Through the dynamic curve changes of diversity, we can capture the microchanges of biological communities, which can be used for multi-scale and cross-scale ecological research.

Combined with the study of Zeta diversity in practical application, some Simons studies have indicated that Zeta diversity, a new diversity research strategy, shows potential in exploring biodiversity, ecological stability and functional research. It can be used to reveal the interrelationships between species and help understand the organizational and functional properties of ecosystems (Simons, A.L.,2023). Mcgeoch's team applied Zeta diversity to areas such as environmental change, biological invasion and ecological restoration to study the effects of human activities and ecosystems (Mcgeoch, M.A.,2017).

2.3.2 Application of exponential regression and power-law regression

Exponential regression and power-law regression are two commonly used nonlinear regression analysis methods used to describe specific relationships between variables. Exponential regression is a nonlinear regression analysis method commonly used to describe the exponential relationship between two variables, and its mathematical model is represented as:

$$y = ab^x \tag{1}$$

Among them, *y* and *x* are variables, and *a* and *b* are parameters for fitting the model. When the data we obtain follows an exponential growth or decay pattern, it can help us better understand the trends in species composition and diversity patterns and reveal their relationship with environmental factors (Rohim et al., 2020).

$$y = ax^b \tag{2}$$

Power law regression is an effective analytical method that helps us better understand the structure and dynamic characteristics of complex systems and reveal their inherent laws and mechanisms when the relationship between data conforms to the power law, that is when the change in one variable is proportional to the change in another variable to a fixed power. In ecological and biodiversity research, these two regression analysis methods have been widely applied to analyze data on species diversity, population distribution, species-area relationships, and other aspects. They are powerful tools and methods for in-depth exploration of the complexity and diversity of ecosystems (Packard et al. 2014).

Using exponential regression and power law regression to help analyze Zeta diversity can reflect much information and influence us. For example, for the attenuation model of shared species, the results of the regression curve will reveal the change in the

number of shared species (zeta diversity) with the increase of the number of samples compared together (zeta order). This change could help researchers understand the decay patterns of shared species in microbial communities in contaminated soils.

At the same time, it can also be used to analyze the similarity and differences of microbial communities, and the change of the exponential regression curve can be used to reveal the ecological differences between different samples or the heterogeneity of contaminated soil. The change of power law regression curve can be used to study whether the microbial community maintains some consistency or common ecological characteristics in contaminated soil (Tettelin et al.,2008).

The comparison results of regression curves can be directly used to reveal the specific effects of pollution on microbial community diversity, sample connectivity and ecological stability. For example, suppose the Zeta diversity regression curve of samples collected from a site decreases faster. In that case, it may indicate that pollution reduces the similarity between microbial communities and the number of shared species, leading to lower connectivity and stability.

2.3.3 Network Analysis

In this study, the IVI influence of microbial communities is used to evaluate the impact of species on the ecological environment. The main factors affecting the intensity of IVI influence are microbial competition, symbiosis and other interactions (Chow, C.E.T.,2014). The interaction and influence between samples will affect the composition region and connection density of network images. Each microbial species in the biological metadata table sampled from the sample is transformed into nodes in the network view, and phi statistics evaluates the intensity of the proportional relationship between different biological groups.

The images formed by Network visualization can directly show the complex structure of the microbial ecosystem, which can be used to help understand the organization and diversity of microbial communities and the influence of different strains in the ecological environment. DeMenezes's team has used network analysis to impact human health monitoring and medical applications (DeMenezes, A.B.,2015).

Through this method, the microbial interactions in the samples were visually presented as a network diagram, showing the patterns of dependence and competition among related species (Barber á n, A.Gen 2012). At the same time, it is also used to speculate the complexity and dynamic changes of microbial communities in the two sample sites. In the same experiment, the biological data collected from two sample sites were classified and summarized, and the relative abundance of microorganisms was calculated according to the metadata. Phi statistics were introduced to measure the relative abundance ratio of two kinds of microorganisms in different samples. Based on these data, each species' importance value index (IVI) is calculated, which mainly depends on the central influence and spread influence of microorganisms. Based on the value of IVI, the microbial network was successfully visualized in the experiment.

2.3.4 Relative abundance

Relative abundance is a fundamental and important measurement indicator in microbial ecology. It describes the proportion of a specific microbial species in the entire microbial community, reflecting its position in the community structure (O'Brien et al. (2011).). The relative abundance can usually be calculated using this formula:

Relative abundance=(number of individuals in a specific substance/total number of microorganisms in the entire sample) × 100%

This experiment used relative abundance to analyze community structure and environmental impact assessment (Santoro et al.,2008). The analysis of the relative abundance of each species in the collected samples can be used to summarize and reveal the composition structure of the microbial community, that is, to identify the dominant species and rare species in the ecological environment. The relative abundance change can reflect the environmental conditions of contaminated soil. Analyzing the changes in relative abundance data of species in different samples can help study how microorganisms respond to and adapt to different degrees of soil pollution.

Meanwhile, the analysis and calculation of relative abundance can reveal the functional species in the ecosystem, namely the microbial species that can interact with PAHs, which is the key to the interaction between microbial communities and the environment and the biodegradation methods in this study. Moreover, the numerical value of relative abundance is also an important perspective for us to evaluate the ecological interference of pollutants on samples. In polluted environments, changes in relative abundance may reveal the impact of interference on microbial communities.

2.3.5 Statistics phi

The statistic phi is a statistic proposed by David Lovell et al. in 2014 to describe the strength of the proportional relationship between two variables (Lovell, D., 2015). The reason for proposing this new statistic is that traditional correlation analysis methods have some limitations when analyzing constituent data. Composition data refers to the overall data composed of several parts, such as relative abundance data in gene expression data. In the composition data, the relative abundance of each component is interdependent, so traditional correlation analysis methods cannot accurately describe the relationship between them.

In contrast, the proportional relationship can provide a more accurate measure of association, so new statistics need to be proposed to describe the strength of the proportional relationship. The statistic phi can more accurately describe the proportional relationship between the constituent data, thereby improving the accuracy and reliability of data analysis. Compared with the traditional correlation analysis method, applying the statistic phi includes gene expression data analysis, microbial composition data analysis, etc. In these applications, the statistic phi can describe the proportional relationship between different genes or microorganisms to understand better their interaction (Paliy & Shankar, 2016). In addition, the statistic phi can also be used to evaluate the fitting degree of different models, thus helping researchers choose the most suitable model.

In practice, the calculation formula for the statistic phi can be expressed as:

$$\phi(\log x, \log y) = \frac{var(\log(\frac{x}{y}))}{var(\log x)} \tag{3}$$

It can also be expressed as:

$$\phi(\log x, \log y) = 1 + \beta^2 - 2\beta|r| \tag{4}$$

Where β is an estimate of the slope used to describe the relationship between the random variables logs and logy, the magnitude of r estimates the strength of the linear relationship between logx and logy. When the value of Phi is closer to zero, the proportional relationship between the variables x and y is stronger. In this study, the variables x and y represent two genes or OTUs in the relative abundance data (Operational et al.).

In this study, the statistic phi was introduced as a more effective analytical tool in network analysis to describe the proportional relationship between genes and to help researchers better understand the interaction between genes. Moreover, others pointed out that the new statistic phi can be used to calculate the weight of edges to better describe the proportional relationship between genes and improve the accuracy of co-expression network analysis (Lovell et al., 2015).

2.3.6 IVI

Integrated Value of Influence (IVI) is an algorithm Abbas Salavaty et al. proposed in 2020 for identifying the most influential nodes in the network. The reason for proposing this algorithm is that existing network centrality measurement methods have some inherent biases and limitations and cannot completely and accurately identify the most influential nodes in the network. Abbas Salavaty's team statistically evaluated 200 real-world and simulated networks, considering multiple network dimensions and integrating the most important and commonly used network centrality measures unbiasedly. The algorithm is a synergistic product of Hubness and spreading values, which can capture all topological dimensions of the network and improve the performance of current tools, so it can deal with the multi-dimensionality and inherent bias of the network, and accurately calculate the influence of each individual in the network (Salavaty et al., 2020).

The IVI algorithm can be expressed as the formula:

$$IVI_i = (Hubness_{score_i})(Spreading_{score_i}) \tag{5}$$

Among them, I represents each node in the network, and the Hubness score is the power index of a node in its surrounding environment, which is calculated by combining the degree centrality and H index centrality of the nodes. Degree centrality refers to the degree of a node, which is the number of edges directly connected to that node. The centrality of the H index considers the degree distribution of neighbouring nodes of a node, that is, the connectivity of the nodes around the node. The hubness score can be used to measure the influence and control of a node in surrounding nodes.

The spreading score is the potential index of a node to spread information in a network, which is calculated by combining four indicators: neighbour connectivity,

clustering coefficient, betweenness centrality, and collective influence. Neighbour connectivity refers to the number of connections between neighbouring nodes of a node, and clustering coefficient refers to the connection density between neighbouring nodes of a node, betweenness centrality refers to the degree to which a node serves as a bridge in the network, and collective influence refers to the overall influence of a node in the network. Therefore, the value of the Spreading score represents the potential of a node to propagate information in the network.

The value of IVI is the product of the Spreading score and Hubness score, which comprehensively considers the propagation potential and power of a node in the network, and is the influence index of a node in the entire network. The higher the IVI, the greater the node's influence in the entire network, that is, the key species that interact with the ecological environment in contaminated soil samples.

3 Results

3.1 Data statistical analysis

3.1.1 Zeta Diversity

This study conducted Zeta diversity analysis on multiple samples collected from two sample points, CH and COV. The obtained images are shown in Figures 3a and 3b. In Figure 3a, the abscissa represents zeta order (number of samples for common comparison), and the ordinate represents zeta diversity (microbial species diversity). The zeta diversity curve of CH sample points is always higher than that of COV sample points. This result indicates that under the same zeta order, there are more shared species in the microbial community of CH sample points, indicating higher microbial diversity. This may suggest that the soil environment of the CH sample site has a more stable microbial community structure.

The curves of both sample points show a clear downward trend. This is because as the number of zeta orders increases, i.e. the number of samples being compared together, the likelihood of finding shared species that all samples exist decreases, leading to a decrease in zeta diversity. This downward trend also reflects the differences and dynamics in species composition within the microbial community. Although the curve of COV sample points is always lower than that of CH sample points, there is a difference in the descent speed between the two. This may indicate a significant difference in species composition between the microbial communities of the two sample sites. This may mean that there may be simpler or more susceptible microbial community composition in the soil of COV sample sites.

It is worth noting that the Zeta diversity values of COV sample points gradually tend to stabilize after the Zeta order is greater than 10, indicating that the microbial community at the location has reached a relatively stable state, and comparing more samples may not significantly increase the number of new species. The PAHs content in soil samples from COV sites is higher than that from CH sites, indicating a stable

trend that may indicate a lower species richness of the microbial community at COV sites.

In contrast, the curve of the CH site still maintained a downward trend in the comparison of 9 samples, indicating that there may still be unsampled species; that is, the species richness of the microbial community at this site may be high. Of course, the CH site's number of samples taken is less than that of the COV site, and the insufficient number of samples taken may also cause this result. At the same time, this difference may also reflect the difference in the interaction and stability of the microbial community between the two sites. Microbial communities at COV sites interact more closely or are more stable, whereas those at CH sites may be more complex and dynamic.

As shown in Figure 4b, the Zeta Ratio curves of the two sample points both showed an upward trend, which indicated that as the number of comparison samples increased, the proportion of shared species to the total number of species also increased. This is because more samples provide more opportunities to discover new shared species. However, the curves of COV sites and CH sites gradually levelled off, the slope of Zeta Ratio gradually decreased, and the diversity of microbial communities became saturated.

Meanwhile, it can be observed that the growth rate of the Zeta Ratio is gradually decreasing. This may indicate that as the number of samples increases, the contribution of new samples to the addition of new shared species gradually decreases. It is worth noting that the Zeta Ratio curve of the CH site is always above the curve of the COV site. This may mean that in the same number of samples, the proportion of shared species in the microbial community of the CH site is higher than that of the COV site. This means that the microbial community of CH sites with less pollution may have higher stability or uniformity compared to heavily polluted areas.

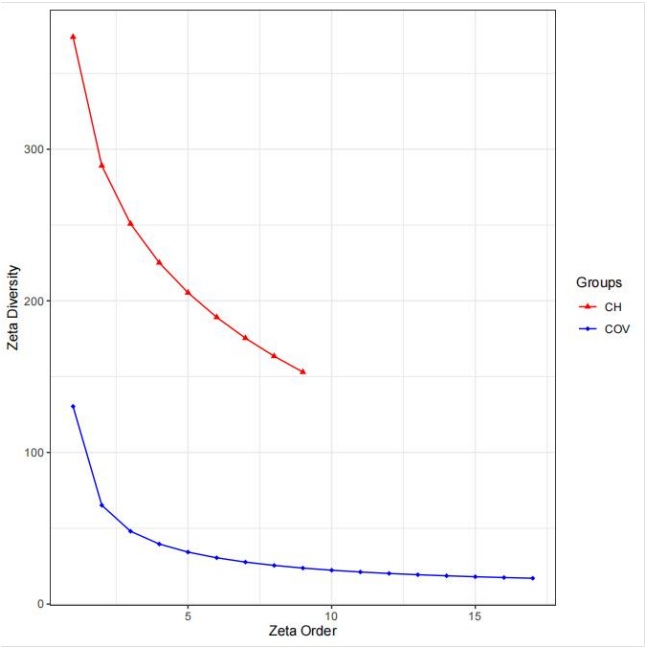
a

	zeta.order	zeta.val	Groups
1	1	130.3529412	COV
2	2	65.15441176	COV
3	3	48.00294118	COV
4	4	39.57773109	COV
5	5	34.25614092	COV
6	6	30.48658694	COV
7	7	27.65811394	COV
8	8	25.46252571	COV
9	9	23.71670095	COV
10	10	22.30131633	COV
11	11	21.13461538	COV
12	12	20.15869425	COV
13	13	19.33151261	COV
14	14	18.62205882	COV
15	15	18.00735294	COV
16	16	17.47058824	COV
17	17	17	COV
18	1	374	CH
19	2	289.1111111	CH
20	3	250.8333333	CH
21	4	225.1269841	CH
22	5	205.3571429	CH
23	6	189.1309524	CH
24	7	175.3611111	CH
25	8	163.4444444	CH
26	9	153	CH

b

	zeta.order	zeta.ratio	Groups
1	1	0.499830776	COV
2	2	0.736756574	COV
3	3	0.824485544	COV
4	4	0.865540797	COV
5	5	0.88995976	COV
6	6	0.907222379	COV
7	7	0.920616849	COV
8	8	0.931435523	COV
9	9	0.940321185	COV
10	10	0.94768466	COV
11	11	0.953823568	COV
12	12	0.958966507	COV
13	13	0.963300659	COV
14	14	0.966990445	COV
15	15	0.970191915	COV
16	16	0.973063973	COV
17	1	0.773024361	CH
18	2	0.867601845	CH
19	3	0.897516216	CH
20	4	0.9121836	CH
21	5	0.920985507	CH
22	6	0.927194142	CH
23	7	0.932044987	CH
24	8	0.936097893	CH

c



d

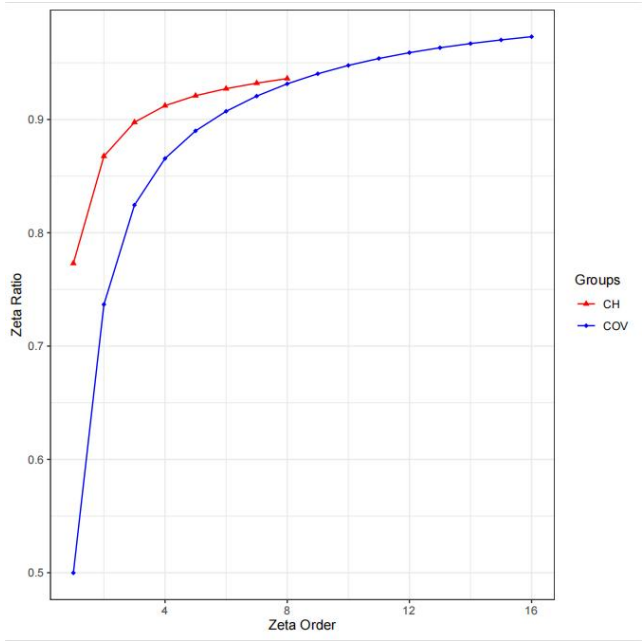


Figure 3: a) Zeta Diversity Table

b) Zeta Ratio Table

c) Zeta Diversity comparison image of COV and CH sample points

d) Zeta Ratio comparison image of COV and CH sample points

3.1.2 Exponential regression and power-law Regression

The biological data collected from soil samples at two sampling points, COV (UK) and CH (Switzerland), were processed and analyzed. The Zeta diversity method of microbial communities was used for evaluation through exponential and power-law regression, and the following results were obtained.

As shown in Figure 4a, in the exponential regression image of the CH sample points, the regression line has a low degree of agreement with the Zeta diversity Zeta sequence data points. However, power-law regression showed a more consistent relationship with Zeta diversity Zeta sequence data, indicating that power-law relationships may be a more suitable model for describing the temporal dynamics of microbial communities at CH sites.

As shown in Figure 4b, in the exponential regression image of COV sample points, the regression line is closely aligned with the Zeta diversity Zeta sequence data points, showing a consistent relationship between all sample points. The fit between the regression line and the Zeta diversity Zeta sequence data points is also relatively high for power-law regression. It can be seen that the microbial community interaction and diversity of COV sites may be affected by higher concentrations of PAHs pollution, which is reflected in both regression models.

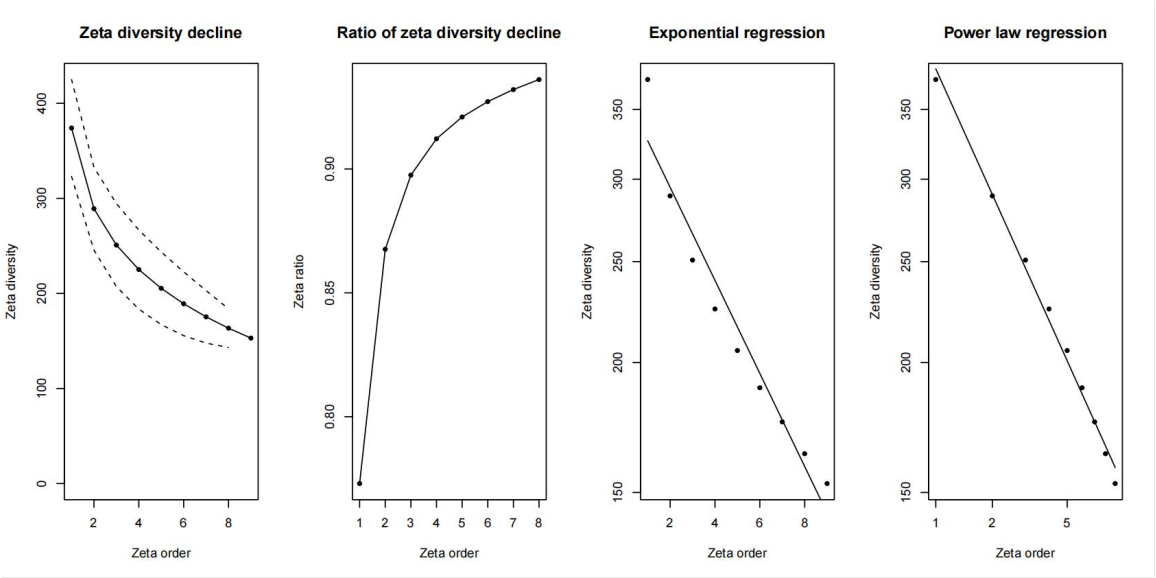
The degree of agreement with two different regression curves can provide different information based on their biological functions. A close exponential regression fit may indicate that the species in the microbial community are relatively stable and less susceptible to external disturbances. If exponential regression no longer fits the data, it means there may be ecological interference or other nonlinear processes.

The power law distribution usually reflects a large number of rare species and a small number of dominant species. This pattern may reveal the structure of microbial

communities, some of which may dominate the community. When the data points closely match the power law regression line, it may indicate that the microbial community has a complex network of interactions.

Based on the analysis of pollutants in the sample, the COV site was significantly affected by PAH pollution, which may limit the diversity and interaction of microbial communities, resulting in both exponential regression and power-law regression closely fitting the data points. The power law regression of the CH site is relatively consistent, which may mean that the microbial community structure of the site contains a large number of rare species and a few dominant species, and the interactions within the community may be more complex.

a



b

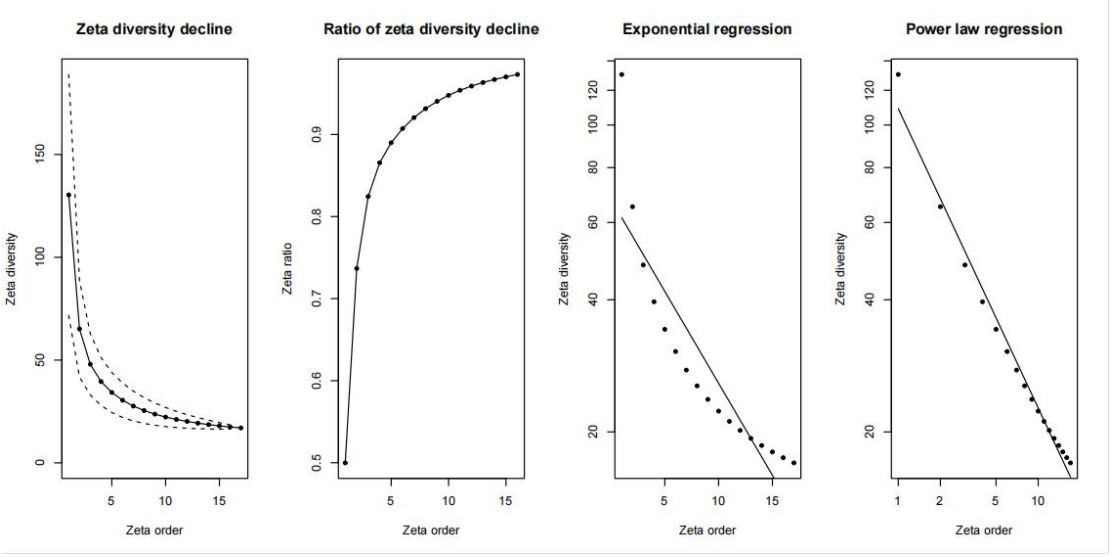


Figure 4: a)Zeta diversity image, Zeta ratio image, exponential regression image, and power-law regression image of CH sample points

b) Zeta diversity image, zeta ratio image, exponential regression image, and power-law regression image of COV sample points

3.1.3 Network Analysis

After calculating the relative abundance and Phi statistics of microbial data in soil samples collected from COV sites in the UK and CH sites in Switzerland, this study used network analysis methods to visualize the microbial network of the two sample points to help analyze the interaction and diversity between microorganisms. As shown in Figure 5, the image shows the key species in the network images of COV and CH sample points. The indicator for evaluating species criticality is based on the Integrated Value of Influence (IVI), which is an indicator used to comprehensively evaluate the importance of species in ecosystems. In this study, it was used as a visual measurement tool. The larger the IVI value of its nodes, the darker the corresponding node colour, indicating a higher evaluation of the importance of this microorganism in soil. In a network graph, a node represents a species, and species of the same category have the same colour. The connections between nodes represent interactions or associations between microorganisms, and the more connections, the more frequent interactions between microorganisms.

As shown in Figure 5a, the node network image of CH sample points presents a well-structured microbial interaction network. In the entire network, the darkest-colored microbial node group represents two major categories of microorganisms, Planctomycota and Firmicutes. Impressively, they exhibited the highest IVI value in this specific soil environment, reaching 100. This not only signifies their dominant position in the entire microbial community but also implies their potential key role in maintaining soil ecological stability and function (Salavaty, A., 2020). At the same time, the three categories of microorganisms, Acidobacterota, Chloroflexi, and Proteobacteria, also show relatively high importance in the figure. Although their IVI values have not reached the highest point, their role in the soil ecosystem cannot be ignored. These microbial categories may be closely related to the nutrient cycling, material transformation, and soil health maintenance functions of the soil, and their presence and activity provide vitality and stability for the ecosystem.

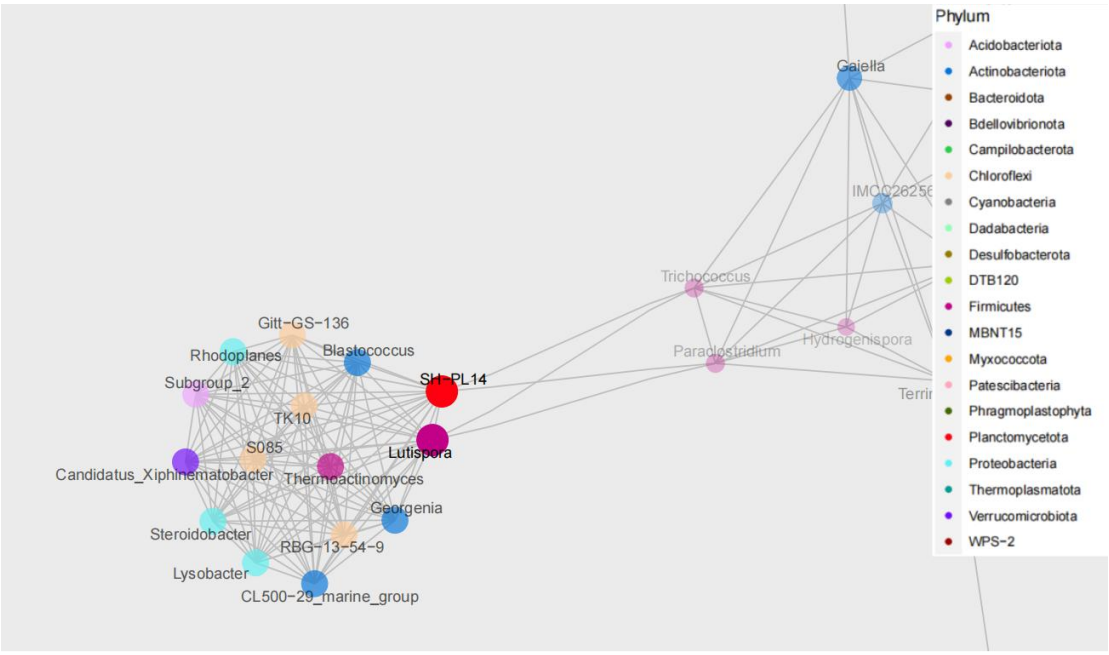
The sample network image of the COV site presents a complex pattern of microbial interaction. In this network, many microbial communities with an IVI value of 100 gather together, and the number of connections between them is enormous. This not only demonstrates their close interaction but also may imply the synergistic effect of these microorganisms in environmental adaptability and resource utilization. Especially the Proteobacteria class, its dominant position in this environment, is particularly evident, occupying over half of the key species positions. Proteobacteria is an extremely diverse phylum of bacteria, particularly important in soil contaminated with polycyclic aromatic hydrocarbons. Many bacteria belonging to Proteobacteria are known to have the ability to decompose polycyclic aromatic hydrocarbons and other complex organic pollutants, such as in Marc Vin In the team research of As, α - Proteobacterium group (Sphingolipids genus) β - Proteobacteria, belonging to the Proteobacterium and Chromobacterium genera, are the dominant species in the degradation of polycyclic aromatic hydrocarbons (Marchand, C., 2017). They can biotransform or completely mineralize these harmful compounds through specific enzyme systems and metabolic pathways, thereby reducing the burden of pollutants in the soil. In addition, some Proteobacteria bacteria form a symbiotic relationship with plant roots, assisting plants in absorbing and transforming harmful substances in the soil, further enhancing the soil's self-cleaning ability. These functions enable Proteobacteria to play a crucial role in contaminated soil, as they not only participate in direct pollutant degradation but may also participate in soil remediation processes together with other microorganisms or plants. Therefore, the dominant position of Proteobacteria in COV site samples may imply the ability of this type of microorganism to respond to high pollution environmental pressures, as well as its potentially important role in bioremediation.

Behind these two images lies the profound impact of polycyclic aromatic hydrocarbon pollution on soil microbial communities. The soil of the COV site selected during the previous sampling process was severely contaminated with polycyclic aromatic hydrocarbons, which may have led to the aggregation of a large number of key

species. Polycyclic aromatic hydrocarbons, as persistent organic pollutants, may cause significant changes in microbial community structure in soil. Some microorganisms, such as Proteobacteria and Acidobacterota, which possess the ability to degrade polycyclic aromatic hydrocarbons, may dominate this environment. On the contrary, the pollution of polycyclic aromatic hydrocarbons in the soil of the CH site is relatively light. However, the significant importance of the Planctomycota and Firmicutes classes still suggests that they may have some association with polycyclic aromatic hydrocarbons. These species may have adapted to this polluted environment and developed mechanisms for interacting with polycyclic aromatic hydrocarbons.

In summary, these two network analysis images not only reveal the soil microbial community structure of the two locations but also provide us with valuable information on the impact of polycyclic aromatic hydrocarbon pollution on soil microorganisms. These findings have important reference value for understanding the role of soil microorganisms in environmental pollution and providing guidance for future soil remediation strategies.

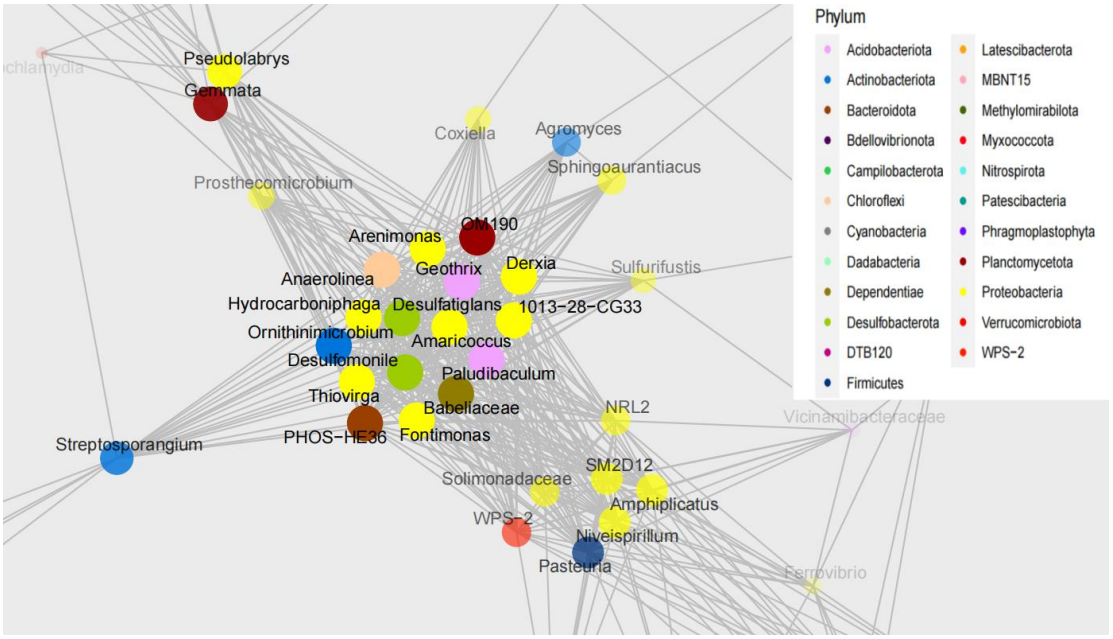
a



b

spreading_score	ivi	hubness_score	Kingdom	Phylum	Class	Order	Family	Genus	label
78.27794341	100	100	Bacteria	Planctomycet	Planctomycet	Planctomycet	Rubinisphaer	SH-PL14	Bacteria;Pl
78.27794341	100	100	Bacteria	Firmicutes	Clostridia	Clostridia	Gracilibacte	Lutispora	Bacteria;Fi
53.28096853	61.47811458	90.27837838	Bacteria	Acidobacteri	Acidobacteri	Subgroup_2	Subgroup_2	Subgroup_2	Bacteria;Ac
53.28096853	61.47811458	90.27837838	Bacteria	Proteobacter	Alphaproteob	Rhizobiales	Xanthobacter	Rhodoplanes	Bacteria;Pr
53.28096853	61.47811458	90.27837838	Bacteria	Chloroflexi	Gitt-GS-136	Gitt-GS-136	Gitt-GS-136	Gitt-GS-136	Bacteria;Ch
53.28096853	61.47811458	90.27837838	Bacteria	Actinobacter	Actinobacter	Micrococcale	Bogoriellace	Georgenia	Bacteria;Ac
53.28096853	61.47811458	90.27837838	Bacteria	Actinobacter	Actinobacter	Frankiales	Geodermatoph	Blastococcus	Bacteria;Ac
53.28096853	61.47811458	90.27837838	Bacteria	Actinobacter	Acidimicrobi	Microtrichal	Ilumatobacte	CL500-29_mar	Bacteria;Ac
53.28096853	61.47811458	90.27837838	Bacteria	Proteobacter	Gammaproteob	Xanthomonada	Xanthomonada	Lysobacter	Bacteria;Pr
53.28096853	61.47811458	90.27837838	Bacteria	Chloroflexi	TK10	TK10	TK10	TK10	Bacteria;Ch
53.28096853	61.47811458	90.27837838	Bacteria	Verrucomicrol	Verrucomicrol	Chthoniobact	Xiphinemato	Candidatus_X	Bacteria;Ve
53.28096853	61.47811458	90.27837838	Bacteria	Chloroflexi	Dehalococcoi	S085	S085	S085	Bacteria;Ch
53.28096853	61.47811458	90.27837838	Bacteria	Firmicutes	Bacilli	Thermoactino	Thermoactino	Thermoactino	Bacteria;Fi
53.28096853	61.47811458	90.27837838	Bacteria	Chloroflexi	Anaerolineae	REG-13-54-9	REG-13-54-9	REG-13-54-9	Bacteria;Ch
53.28096853	61.47811458	90.27837838	Bacteria	Proteobacter	Gammaproteob	Steroidobact	Steroidobact	Steroidobact	Bacteria;Pr
100	51.5820383	39.88648649	Bacteria	Actinobacter	Thermoleophi	Gaiellales	Gaiellaceae	Gaiella	Bacteria;Ac

c



d

spreading_score	ivi	hubness_score	Kingdom	Phylum	Class	Order	Family	Genus	label
82.26893328	100	100	Bacteria	Desulfobacte	Desulfomonil	Desulfomonil	Desulfomonil	Desulfomonil	Bacteria;Des
82.26893328	100	100	Bacteria	Acidobacteri	Holophagae	Holophagales	Holophagacea	Geothrix	Bacteria;Aci
82.26893328	100	100	Bacteria	Chloroflexi	Anaerolineae	Anaerolineal	Anaerolineac	Anaerolinea	Bacteria;Chl
82.26893328	100	100	Bacteria	Actinobacter	Actinobacter	Micrococcale	Intrasporang	Ornithinimic	Bacteria;Act
82.26893328	100	100	Bacteria	Proteobacter	Gammaproteob	Salinisphaer	Solimonadace	Fontimonas	Bacteria;Pro
82.26893328	100	100	Bacteria	Dependentiae	Babeliae	Babeliales	Babeliaceae	Babeliaceae	Bacteria;Dep
82.26893328	100	100	Bacteria	Acidobacteri	Acidobacteri	Acidobacteri	Acidobacteri	Paludibaculu	Bacteria;Aci
82.26893328	100	100	Bacteria	Desulfobacte	Desulfobacte	Desulfatigla	Desulfatigla	Desulfatigla	Bacteria;Des
82.26893328	100	100	Bacteria	Bacteroidota	Ignavibacter	Ignavibacter	PHOS-HE36	PHOS-HE36	Bacteria;Bac
82.26893328	100	100	Bacteria	Proteobacter	Gammaproteob	Burkholderia	Alcaligenace	Derxia	Bacteria;Pro
82.26893328	100	100	Bacteria	Proteobacter	Gammaproteob	1013-28-CG33	1013-28-CG33	1013-28-CG33	Bacteria;Pro
82.26893328	100	100	Bacteria	Proteobacter	Gammaproteob	Salinisphaer	Solimonadace	Hydrocarboni	Bacteria;Pro
82.26893328	100	100	Bacteria	Proteobacter	Alphaproteob	Rhodobactera	Rhodobactera	Amaricoccus	Bacteria;Pro
82.26893328	100	100	Bacteria	Proteobacter	Gammaproteob	Xanthomonada	Xanthomonada	Arenimonas	Bacteria;Pro
82.26893328	100	100	Bacteria	Planctomycet	OM190	OM190	OM190	OM190	Bacteria;Pla
82.26893328	100	100	Bacteria	Proteobacter	Gammaproteob	Halothiobaci	Halothiobaci	Thiovirga	Bacteria;Pro
96.34684233	90.27775244	76.86746335	Bacteria	Proteobacter	Alphaproteob	Rhizobiales	Xanthobacter	Pseudolabrys	Bacteria;Pro
96.34684233	90.27775244	76.86746335	Bacteria	Planctomycet	Planctomycet	Gemmatales	Gemmataceae	Gemmata	Bacteria;Pla
100	81.44264218	66.7052963	Bacteria	Actinobacter	Actinobacter	Streptospora	Streptospora	Streptospora	Bacteria;Act
61.32482627	71.27959429	95.62799401	Bacteria	Proteobacter	Alphaproteob	Rickettsiale	SM2D12	SM2D12	Bacteria;Pro
61.32482627	71.27959429	95.62799401	Bacteria	Proteobacter	Alphaproteob	Caulobactera	Parvularcula	Amphiplicatu	Bacteria;Pro
61.32482627	71.27959429	95.62799401	Bacteria	Firmicutes	Bacilli	Thermoactino	Thermoactino	Pasteuria	Bacteria;Fir
61.32482627	71.27959429	95.62799401	Bacteria	Proteobacter	Alphaproteob	Azospirillal	Azospirillac	Niveispirill	Bacteria;Pro
52.87901696	57.08262542	88.72886124	Bacteria	WPS-2	WPS-2	WPS-2	WPS-2	WPS-2	Bacteria;WPS
52.87901696	57.08262542	88.72886124	Bacteria	Proteobacter	Alphaproteob	NRL2	NRL2	NRL2	Bacteria;Pro
52.87901696	57.08262542	88.72886124	Bacteria	Proteobacter	Gammaproteob	Salinisphaer	Solimonadace	Solimonadace	Bacteria;Pro
72.08377695	53.94454739	61.13999587	Bacteria	Actinobacter	Actinobacter	Micrococcale	Microbacteri	Agromyces	Bacteria;Act
72.08377695	53.94454739	61.13999587	Bacteria	Proteobacter	Alphaproteob	Sphingomonad	Sphingomonad	Sphingoauran	Bacteria;Pro

Figure 5: a) Key species in the network image of CH sample points based on IVI values.

b) CH sample points are based on the IVI value table.

c) The key species in the network image of Cov sample points based on IVI values.
d) The COV sample points are based on the IVI numerical table. The depth of the color represents the size of the IVI value, and nodes of the same color represent species of the same category.

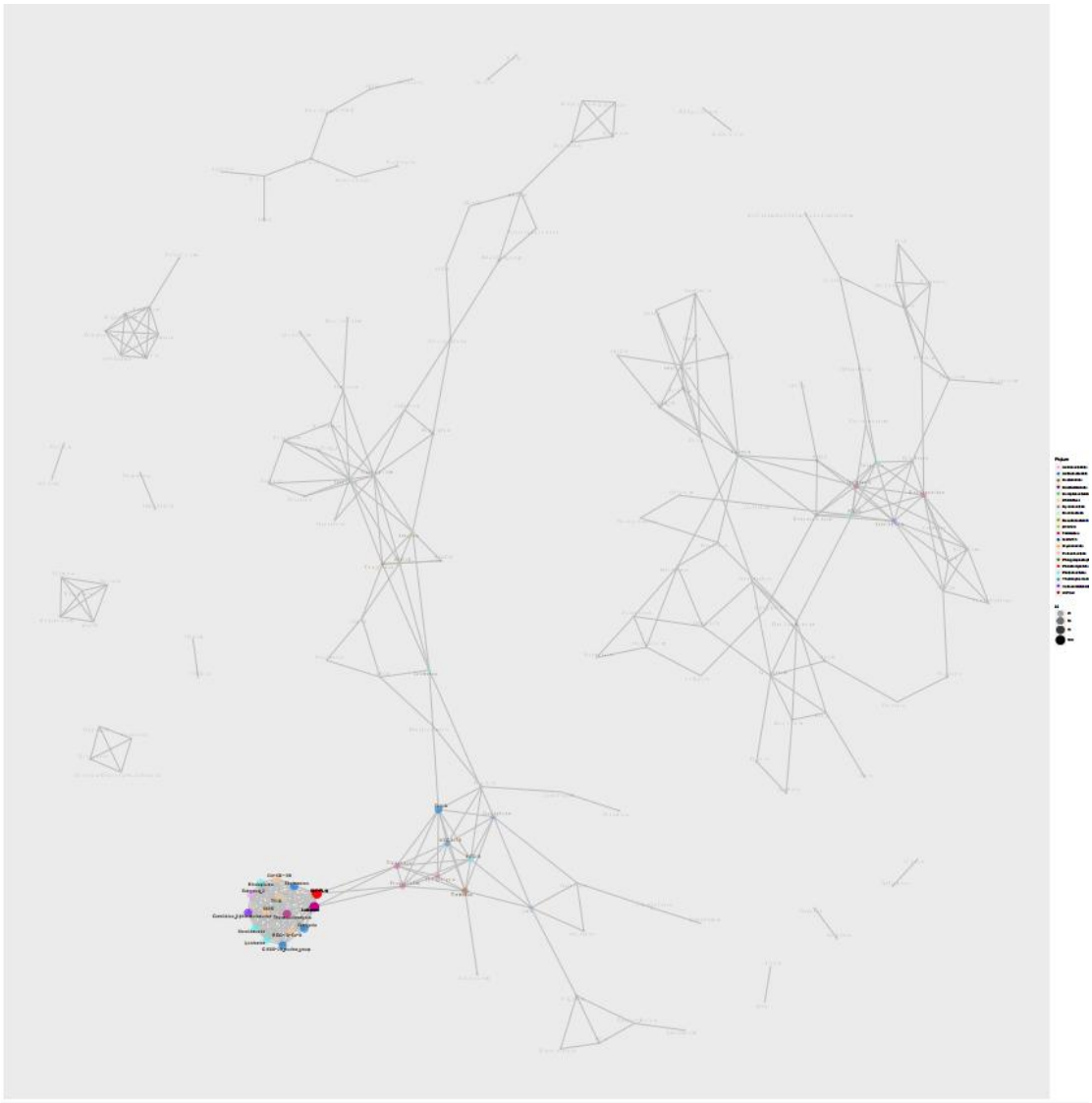
In addition to analyzing the IVI values of each species, the interaction patterns of microbial communities at these two sites also reveal in-depth information about soil health and function. As shown in Figures 6, a, and b, the close interactions between a large number of key species at the COV site may imply a highly collaborative ecological network. For example, Proteobacteria, Desulfobacterota, and Dependencies may be involved in the initial decomposition of polycyclic aromatic hydrocarbons, converting them into simpler organic compounds. Firmicutes and Actinobacteriota may further transform these organic compounds, ultimately mineralizing them into harmless substances such as carbon dioxide and water. In addition, Acidobacterota and Chloroflexi may produce certain bioactive substances, such as enzymes and biosurfactants, to enhance the degradation ability of other microorganisms, thereby accelerating the degradation process of pollutants.

Bacteroidota and Myxococota may form a symbiotic relationship with plant roots, helping plants obtain essential nutrients while also obtaining organic carbon from plants. This symbiotic relationship can enhance the structural stability of the soil, promote the retention of water and nutrients, and provide a more stable living environment for microorganisms. In polluted soil, this interdependence may be more pronounced because, in such environments, resources may be more limited, and there may be more ecological pressure.

In contrast, the microbial interaction mode of the CH site is relatively simple. The IVI values of Planctomycota and Firmicutes in this environment are both 100, indicating their significant importance in this soil environment. In addition, the Acidobacterota, Chloroflexi, and Proteobacteria classes also show relatively high importance. Except

for key species, there are fewer interactions between certain species, possibly because they have sufficient resources to support their growth in low-polluting soil without needing to form a close interaction network with other microorganisms. For example, Actinobacteriota and Thermoplasmatota may have independent functions in decomposing organic matter without relying on the help of other microorganisms. However, this simplified network may also mean that the functionality and stability of the soil are limited, as the lack of diversity and interaction may lead to a decrease in the soil's resistance to external pressure.

a



b

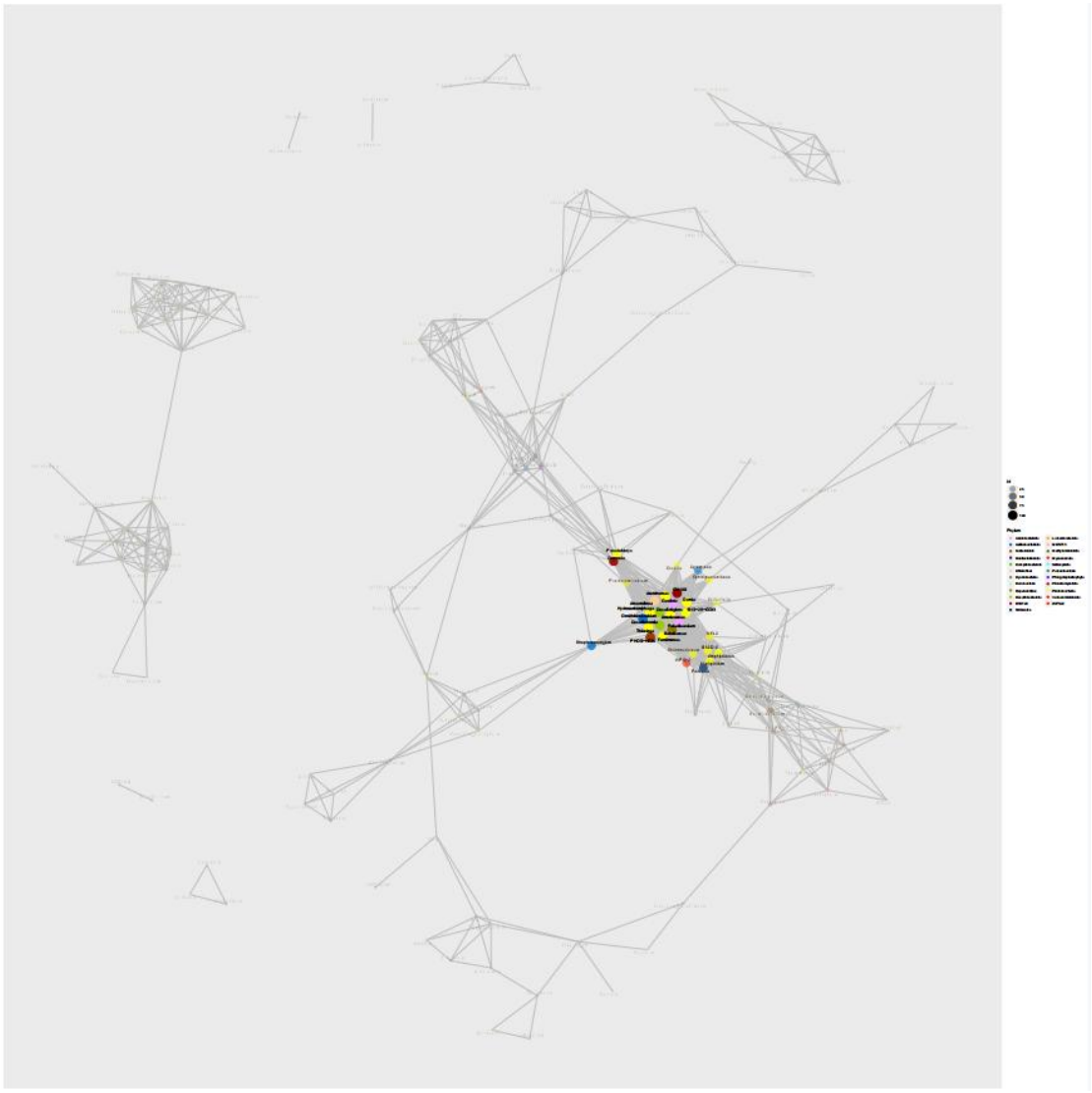
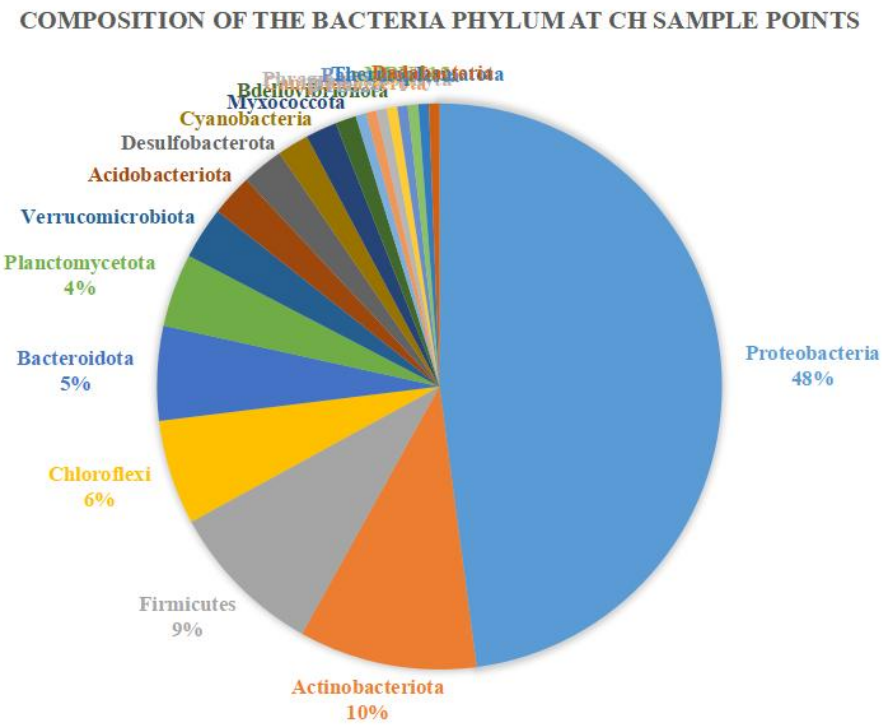


Figure 6: a) Overall network image of CH sample points based on IVI values

b) Overall network image of COV sample points based on IVI values

a



b

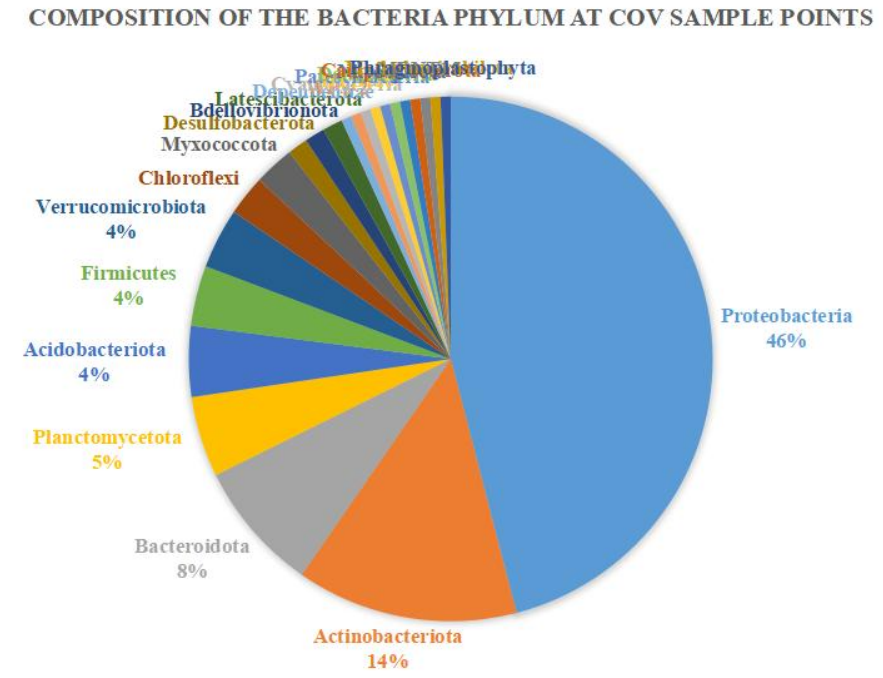


Figure 7: a) Composition of bacterial phylum at CH sampling point

b) Composition of bacterial phylum at COV sampling point

As shown in Figure 7a, the proportion of Proteobacteria bacteria is 48%, which is the highest proportion of bacteria in the CH sample point. Actinobacteriota and

Firmicutes bacteria are second, accounting for 10% and 9%, respectively. However, according to the IVI values of this sample point in Figure 6b, the species that play a crucial role in this soil sample are not Proteobacteria, but Planctomycota, Actinobacteriota, and Chloroflexi. This indicates that relying solely on species richness or proportion may mislead our judgment of the importance of a particular species in the ecosystem.

As shown in Figure 7b, in the COV (heavily polluted) sample points, the proportion of Proteobacteria bacteria is 46%, which is the highest proportion of bacteria in the COV sample points. Actinobacteriota and Bacteroidota bacteria are second, accounting for 14% and 8%, respectively. Meanwhile, according to the IVI values of the sample point in Figure 6d, the species that play a key role in the soil sample are still Proteobacteria, Actinobacteriota, and Bacteroidota categories of bacteria. This may mean that certain bacterial species can better adapt and survive under heavier pollution conditions and may play a more critical role in ecological processes.

4 Discussion

4.1 Temporal dynamics of microbial communities in soil samples from COV and CH sample points

The image results of zeta diversity obtained from experiments can be used to analyze and study the temporal dynamics of microbial communities. The results showed that with the increase of the number of samples for common comparison, i.e. the increase of zeta order, the zeta diversity of both sample points showed a significant downward trend. This is because as the number of samples increases, the likelihood of finding shared species that all samples exist decreases, leading to a decrease in zeta diversity.

The different downward trends of the two curves also reflect the differences and dynamics in the species composition within the microbial communities in the two soil samples. Even though the zeta diversity of COV sample points is always lower than that of CH sample points, there is a difference in the rate of decline between the two. This means that the microbial communities of the two sample points may have significant differences in species composition due to the influence of pollutants such as polycyclic aromatic hydrocarbons. This means that in areas with severe polycyclic aromatic hydrocarbon pollution, the number of species is relatively small, and most species are rare. The diversity curve of COV sample points gradually shows a stable trend, indicating that in this sample, most species are rare, with only a few species widely distributed. This minority species may play important ecological functions under polluted environmental conditions, and the species composition of the community tends to be stable.

Secondly, based on the comparison of zeta ratios between the two sample points, it can be seen that the proportion of shared species in the microbial community of the CH site is higher than that of the COV site within a limited zeta order. However, when the zeta order is greater than 8, the proportion of shared species in the two sample points is close, and the zeta ratio curve of the CH site tends to stabilize from an

upward trend. This means that when continuing to increase the comparison of sample numbers, The proportion of shared species in the microbial community of the CH site may be surpassed by the COV site, which suggests that this COV site may already have higher stability and uniformity while the composition and interaction relationship of CH site's biological community is relatively dynamic. At the same time, both power law regression curves are in a declining state, which means that both samples are under environmental pressure from pollutants.

4.2 Changes in driving factors and interactions within communities

Based on the analysis of Zeta diversity, it is possible to explore the changes in driving factors and interactions within microbial communities as soil pollution increases the pressure on microbial communities.

At the time of sample collection, it was detected that the soil of the COV site had a high content of polycyclic aromatic hydrocarbons and other pollutants, while the soil pollution of the CH site was relatively light. Therefore, we can believe that the data from the CH site can represent the ecological environment when soil pollution has a weak impact on microbial communities, while the data from the COV site can represent the ecological environment when soil pollution has a strong impact on microbial communities. As soil pollution intensifies, the driving factors of microbial communities have shifted. At the CH site, the driving factors of microbial communities are mainly environmental factors such as soil pH and organic matter content, while at the COV site, the driving factors of microbial communities are mainly the presence of pollutants such as polycyclic aromatic hydrocarbons. This indicates that the driving factor of microbial communities has shifted from environmental resource factors to pollutant factors (Alexander et al., 1999).

Secondly, the interactions within the microbial community have also changed with the aggravation of soil pollution. In the environment of CH sites, there is relatively less competition for resources among microorganisms, and the symbiotic relationship and mutual promotion between species are more significant. Therefore, the interactions within the microbial community are mainly positive; that is, the symbiotic relationship and mutual promotion are more significant. At COV sites, the interactions within the microbial community are more complex, with both positive and negative interactions. For example, Proteobacterium is an important degrading bacterium for polycyclic aromatic hydrocarbons, and they are numerous in COV samples, most of which are key functional species. However, there are also some competitive and

antagonistic relationships between degrading bacteria. In addition, multiple pollutants are present in the soil of COV sites, and different microorganisms may have different degradation capabilities and adaptability to different pollutants. This indicates that as soil pollution intensifies, the interactions within microbial communities become more complex and diverse, and there may be some competitive and antagonistic relationships, leading to more complex and diverse interactions within microbial communities.

4.3 Key species involved in the biodegradation process of soil pollution

This project is focused on an examination of natural ecosystems and core microbiota under various levels of pollution. The results of zeta diversity and network analysis can be used to investigate the composition and interaction relationships of microbial communities, showing the role and mechanism of microorganisms in the biodegradation process.

The Zeta diversity of bacterial taxa such as Plantomycetota, Firmicutes, and Actinobacteriota in the microbial community at the CH site is relatively high, according to the chart results based on IVI values. The Plantomycetota genus, on the other hand, is a crucial participant in the nitrification and denitrification processes and plays an important role in the CH sample points. However, proof for its specific role in PAH breakdown is currently lacking.

Firmicutes and Actinobacteria, among other bacterial genera, have multiple members capable of producing a series of enzymes to catalyze the oxidation of polycyclic aromatic hydrocarbons. These bacterial genera may play an important role in the degradation process of polycyclic aromatic hydrocarbon (Huesemann et al., 1995). At the COV site, the Zeta diversity of Proteobacteria and Acidobacteriota genera in the microbial community is relatively high. Proteobacteria genera have multiple species of bacteria, such as Pseudobacteria, Alcaliginaceae, and Desulfatiguans, which have been proven to be able to utilize PAHs as carbon and energy sources. These bacterial genera play an important role in the degradation process of polycyclic aromatic hydrocarbons.

Through the network analysis of the image, it can be concluded that the interaction between Proteus, Bacillus and other genera in the microbial community at CH site is relatively loose, and these genera may play a relatively independent role in the

degradation of PAHs. There is a close interaction among other genera of bacteria, which may play a synergistic role in the degradation of PAHs. In the COV site, there is a close interaction between Proteus and actinomycetes in the microbial community, which may play a synergistic effect on the degradation of PAHs.

Based on the above analysis results, it can be preliminarily speculated that Proteobacteria and Acidobacteriota may be one of the key degrading bacterial communities in the biodegradation process of soil pollution caused by polycyclic aromatic hydrocarbon lipophilic organic pollutants. At the same time, Firmicutes and Actinobacteriota may also play an important role in the degradation process. These bacterial genera may participate in the degradation process of polycyclic aromatic hydrocarbons through synergistic or relatively independent effects, playing a key role.

4.4 The richness of species and their crucial role in ecosystems

After studying soil samples with different levels of CH and COV pollution, we conducted a statistical comparison of the structure and richness of microbial communities in the soil. Firstly, we found that the richness of species is not always consistent with their critical role in the ecosystem. For example, although the proportion of Proteobacteria bacteria in the CH sample point is 48%, which is the highest proportion of bacteria in the CH sample point, according to the IVI values in Figure 6b, they are not the dominant species in the soil sample. Therefore, speculating solely on the richness or proportion of species may mislead our judgment of the importance of a particular species in the ecosystem.

In addition, in COV sample points (heavily contaminated with PAHs), bacterial categories such as Proteobacteria, Actinobacteriota, and Bacteroidota have not only abundant numbers but also play key core roles in the ecosystem. This means that under heavier pollution conditions, certain bacterial species can better adapt and survive, which may depend on the interactions in the ecological network. Some bacterial categories that dominate in quantity may interact less with other organisms, so their influence in ecological networks is relatively small. Other bacterial categories with smaller numbers may have a more central position in the ecological network due to their interactions with more organisms.

5 Conclusions and Future work

This study utilizes zeta diversity and network analysis methods to explore the composition, interaction relationships, and temporal dynamics of microbial communities, revealing the interactions and mechanisms of microorganisms in the biodegradation process of organic pollutants such as polycyclic aromatic hydrocarbons.

Through zeta diversity analysis, we found differences in the composition of microbial communities in different sites, in which bacteria such as Planctomycetota, Firmicutes and Actinobacteria played an essential role in the microbial community in CH sites. In contrast, bacteria such as Proteobacteria and Acidobacteriota played an essential role in the microbial community in COV sites. Most of these bacteria have metabolic mechanisms for dealing with PAHs and other pollutants. Therefore, Proteobacteria and Acidobacteriota are potential species to study the future biodegradation methods of contaminated soils.

Through network analysis, we found that there was a loose interaction between bacteria and other genera in the CH site, Planctomycetota, Firmicutes and Actinobacteria, while in the COV site, the interaction between Proteobacteria and other genera was closer, and there was a more complex interaction.

In addition, we also found that microbial communities diversity and temporal dynamics were related to the concentration and distribution of organic pollutants such as polycyclic aromatic hydrocarbons. In the COV site, the diversity of the microbial community is low. However, the composition and interaction of microbial communities tend to be stable gradually, which may be related to the adaptability of microbial species caused by the concentration of PAHs in the site. In the CH site, the diversity of the microbial community is higher. However, the composition and interaction of the microbial community are more dynamic, and the composition of the

microbial community could be more stable in the limited samples, which may be related to the uneven concentration and distribution of PAHs in the site.

Based on the above analysis results, microbial communities' composition, interaction and diversity are closely related to the existence and distribution of PAHs and other organic pollutants. Future research can further explore the time-dynamic changes of microbial communities and the relationship between microbial communities and environmental factors to understand better the role and mechanism of microorganisms in the biodegradation of organic pollutants. In-depth study of the function and metabolic pathway of the microbial community in order to improve the biodegradation efficiency of organic pollutants. The research focus can also be extended to the biodegradation process of other organic pollutants, such as pesticides, petroleum hydrocarbons, etc., to expand the research field and provide more comprehensive and effective solutions for environmental pollution control.

Reference

1. Abdel-Shafy, H. I., & Mansour, M. S. (2016). A review on polycyclic aromatic hydrocarbons: source, environmental impact, effect on human health and remediation. *Egyptian journal of petroleum, 25*(1), 107-123.
2. Abellan-Schneyder, I., Matchado, M. S., Reitmeier, S., Sommer, A., Sewald, Z., Baumbach, J., . . . Neuhaus, K. (2021). Primer, pipelines, parameters: issues in 16S rRNA gene sequencing. Msphere, 6(1), 10.1128/msphere. 01202-01220.
3. Abhilash, P., Dubey, R. K., Tripathi, V., Srivastava, P., Verma, J. P., & Singh, H. (2013). Remediation and management of POPs-contaminated soils in a warming climate: challenges and perspectives. Environmental Science and Pollution Research, 20, 5879-5885.
4. Ahmad, A. A., Muhammad, I., Shah, T., Kalwar, Q., Zhang, J., Liang, Z., . . . Zhi, D. (2020). Remediation methods of crude oil contaminated soil. World Journal of Agriculture and Soil Science, 4(3), 8.
5. Alexander, M. (1999). Biodegradation and bioremediation: Gulf Professional Publishing.
6. Ayangbenro, A. S., & Babalola, O. O. (2017). A new strategy for heavy metal polluted environments: a review of microbial biosorbents. International journal of environmental research and public health, 14(1), 94.
7. Barberán, A., Bates, S. T., Casamayor, E. O., & Fierer, N. (2012). Using network analysis to explore co-occurrence patterns in soil microbial communities. The ISME journal, 6(2), 343-351.

8. Bowman, B. A., & Kwon, D. S. (2016). Efficient nucleic acid extraction and 16S rRNA gene sequencing for bacterial community characterization. JoVE (Journal of Visualized Experiments)(110), e53939.

9. Choi, H., Harrison, R., Komulainen, H., & Saborit, J. M. D. (2010). Polycyclic aromatic hydrocarbons. In WHO guidelines for indoor air quality: selected pollutants: World Health Organization.

10. Chow, C.-E. T., Kim, D. Y., Sachdeva, R., Caron, D. A., & Fuhrman, J. A. (2014). Top-down controls on bacterial community structure: microbial network analysis of bacteria, T4-like viruses and protists. The ISME journal, 8(4), 816-829.

11. De Menezes, A. B., Prendergast-Miller, M. T., Richardson, A. E., Toscas, P., Farrell, M., Macdonald, L. M., . . . Thrall, P. H. (2015). Network analysis reveals that bacteria and fungi form modules that correlate independently with soil parameters. Environmental microbiology, 17(8), 2677-2689.

12. França, L. T., Carrilho, E., & Kist, T. B. (2002). A review of DNA sequencing techniques. Quarterly reviews of biophysics, 35(2), 169-200.

13. Freeman, D. J., & Cattell, F. C. (1990). Woodburning as a source of atmospheric polycyclic aromatic hydrocarbons. Environmental science & technology, 24(10), 1581-1585.

14. Gan, S., Lau, E., & Ng, H. K. (2009). Remediation of soils contaminated with polycyclic aromatic hydrocarbons (PAHs). Journal of hazardous materials, 172(2-3), 532-549.

15. 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.

16. Ghosal, D., Ghosh, S., Dutta, T. K., & Ahn, Y. (2016). Current state of knowledge in microbial degradation of polycyclic aromatic hydrocarbons (PAHs): a review. Frontiers in Microbiology, 1369.

17. Huesemann, M. H. (1995). Predictive model for estimating the extent of petroleum hydrocarbon biodegradation in contaminated soils. Environmental science & technology, 29(1), 7-18.

18. Hui, C., & McGeoch, M. A. (2014). Zeta diversity as a concept and metric that unifies incidence-based biodiversity patterns. The American Naturalist, 184(5), 684-694.

19. Jackson, M. A., Bell, J. T., Spector, T. D., & Steves, C. J. (2016). A heritability-based comparison of methods used to cluster 16S rRNA gene sequences into operational taxonomic units. PeerJ, 4, e2341.

20. Janda, J. M., & Abbott, S. L. (2007). 16S rRNA gene sequencing for bacterial identification in the diagnostic laboratory: pluses, perils, and pitfalls. Journal of clinical microbiology, 45(9), 2761-2764.

21. Kuppusamy, S., Thavamani, P., Venkateswarlu, K., Lee, Y. B., Naidu, R., & Megharaj, M. (2017). Remediation approaches for polycyclic aromatic

hydrocarbons (PAHs) contaminated soils: Technological constraints, emerging trends and future directions. Chemosphere, 168, 944-968.

22. Legendre, P. (2008). Studying beta diversity: ecological variation partitioning by multiple regression and canonical analysis. Journal of plant ecology, 1(1), 3-8.

23. Lim, M. W., Von Lau, E., & Poh, P. E. (2016). A comprehensive guide of remediation technologies for oil contaminated soil—Present works and future directions. Marine pollution bulletin, 109(1), 14-45.

24. Liu, L., Li, W., Song, W., & Guo, M. (2018). Remediation techniques for heavy metal-contaminated soils: Principles and applicability. Science of the total environment, 633, 206-219.

25. Lladó Fernández, S., Větrovský, T., & Baldrian, P. (2019). The concept of operational taxonomic units revisited: genomes of bacteria that are regarded as closely related are often highly dissimilar. Folia microbiologica, 64, 19-23.

26. Lovell, D., Pawlowsky-Glahn, V., Egozcue, J. J., Marguerat, S., & Bähler, J. (2015). Proportionality: a valid alternative to correlation for relative data. PLoS computational biology, 11(3), e1004075.

27. Marchand, C., St-Arnaud, M., Hogland, W., Bell, T. H., & Hijri, M. (2017). Petroleum biodegradation capacity of bacteria and fungi isolated from petroleum-contaminated soil. International Biodeterioration & Biodegradation, 116, 48-57.

28. Margesin, R., & Schinner, F. (2005). Manual for soil analysis-monitoring and assessing soil bioremediation (Vol. 5): Springer Science & Business Media.

29. Mcgeoch, M. A., Latombe, G., Andrew, N. R., Nakagawa, S., Nipperess, D. A., Roige, M., . . . Thomas, T. (2017). The application of zeta diversity as a continuous measure of compositional change in ecology. Biorxiv, 216580.

30. Mirsal, I. A. (2008). Soil pollution: Springer.

31. Mirsal, I. A., & Mirsal, I. A. (2008). Sources of soil pollution. Soil Pollution: Origin, Monitoring & Remediation, 137-173.

32. Mishra, R. K., Mohammad, N., & Roychoudhury, N. (2016). Soil pollution: Causes, effects and control. Van Sangyan, 3(1), 1-14.

33. Mojiri, A., Zhou, J. L., Ohashi, A., Ozaki, N., & Kindaichi, T. (2019). Comprehensive review of polycyclic aromatic hydrocarbons in water sources, their effects and treatments. Science of the total environment, 696, 133971.

34. Nguyen, N.-P., Warnow, T., Pop, M., & White, B. (2016). A perspective on 16S rRNA operational taxonomic unit clustering using sequence similarity. NPJ biofilms and microbiomes, 2(1), 1-8.

35. Packard, G. C. (2014). On the use of log-transformation versus nonlinear regression for analyzing biological power laws. Biological Journal of the Linnean Society, 113(4), 1167-1178.

36. Paliy, O., & Shankar, V. (2016). Application of multivariate statistical techniques in microbial ecology. Molecular ecology, 25(5), 1032-1057.

37. Park, S. T., & Kim, J. (2016). Trends in next-generation sequencing and a new era for whole genome sequencing. International neurourology journal, 20(Suppl 2), S76.

38. Patel, A. B., Shaikh, S., Jain, K. R., Desai, C., & Madamwar, D. (2020). Polycyclic aromatic hydrocarbons: sources, toxicity, and remediation approaches. Frontiers in Microbiology, 11, 562813.

39. Preheim, S. P., Perrotta, A. R., Martin-Platero, A. M., Gupta, A., & Alm, E. J. (2013). Distribution-based clustering: using ecology to refine the operational taxonomic unit. Applied and environmental microbiology, 79(21), 6593-6603.

40. Ravindra, K., Sokhi, R., & Van Grieken, R. (2008). Atmospheric polycyclic aromatic hydrocarbons: source attribution, emission factors and regulation. Atmospheric environment, 42(13), 2895-2921.

41. Rohim, R. A. A., Ahmad, W. M. A. W., Ismail, N. H., Ghazali, F. M. M., & Alam, M. K. (2020). Modeling the growth of bacteria streptococcus sobrinus using exponential regression. Pesquisa Brasileira em Odontopediatria e Clínica Integrada, 20, e5380.

42. Salavaty, A., Ramialison, M., & Currie, P. D. (2020). Integrated value of influence: an integrative method for the identification of the most influential nodes within networks. Patterns, 1(5).

43. Samanta, S. K., Singh, O. V., & Jain, R. K. (2002). Polycyclic aromatic hydrocarbons: environmental pollution and bioremediation. TRENDS in Biotechnology, 20(6), 243-248.

44. Santoro, A. E., Francis, C. A., De Sieyes, N. R., & Boehm, A. B. (2008). Shifts in the relative abundance of ammonia-oxidizing bacteria and archaea across

physicochemical gradients in a subterranean estuary. Environmental microbiology, 10(4), 1068-1079.

45. Simons, A. L., Theroux, S., Osborne, M., Nuzhdin, S., Mazor, R., & Steele, J. (2023). Zeta diversity patterns in metabarcoded lotic algal assemblages as a tool for bioassessment. Ecological Applications, 33(3), e2812.

46. Statnikov, A., Henaff, M., Narendra, V., Konganti, K., Li, Z., Yang, L., . . . Alekseyenko, A. V. (2013). A comprehensive evaluation of multicategory classification methods for microbiomic data. Microbiome, 1(1), 1-12.

47. Sun, S., Sidhu, V., Rong, Y., & Zheng, Y. (2018). Pesticide pollution in agricultural soils and sustainable remediation methods: a review. Current Pollution Reports, 4, 240-250.

48. Tettelin, H., Riley, D., Cattuto, C., & Medini, D. (2008). Comparative genomics: the bacterial pan-genome. Current opinion in microbiology, 11(5), 472-477.

49. Thukral, A. K. (2017). A review on measurement of Alpha diversity in biology. Agricultural Research Journal, 54(1).

50. Wang, D., Yang, M., Jia, H., Zhou, L., & Li, Y. (2008). Seasonal variation of polycyclic aromatic hydrocarbons in soil and air of Dalian areas, China: an assessment of soil–air exchange. Journal of Environmental Monitoring, 10(9), 1076-1083.

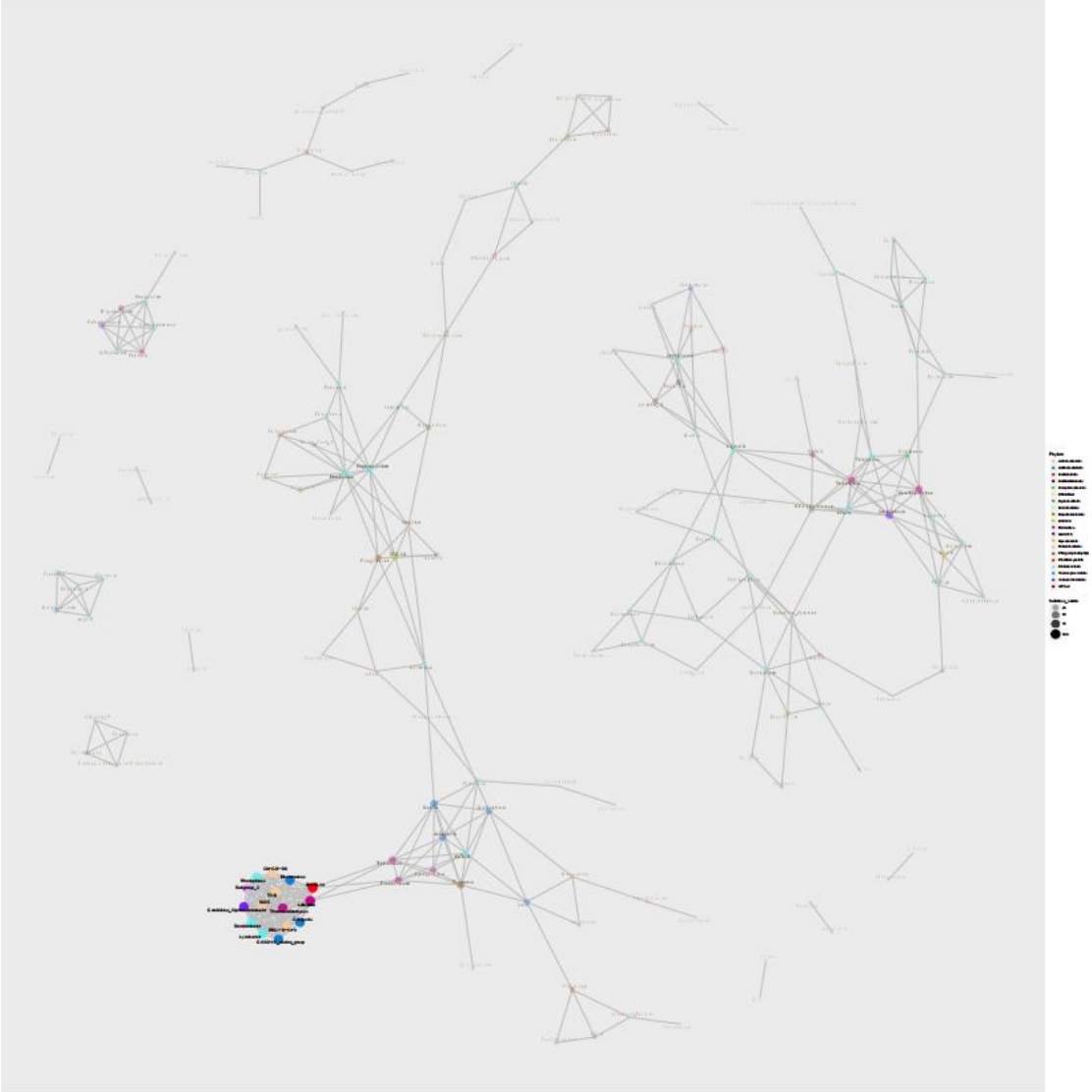
51. Yoon, S.-H., Ha, S.-M., Kwon, S., Lim, J., Kim, Y., Seo, H., & Chun, J. (2017). Introducing EzBioCloud: a taxonomically united database of 16S rRNA gene

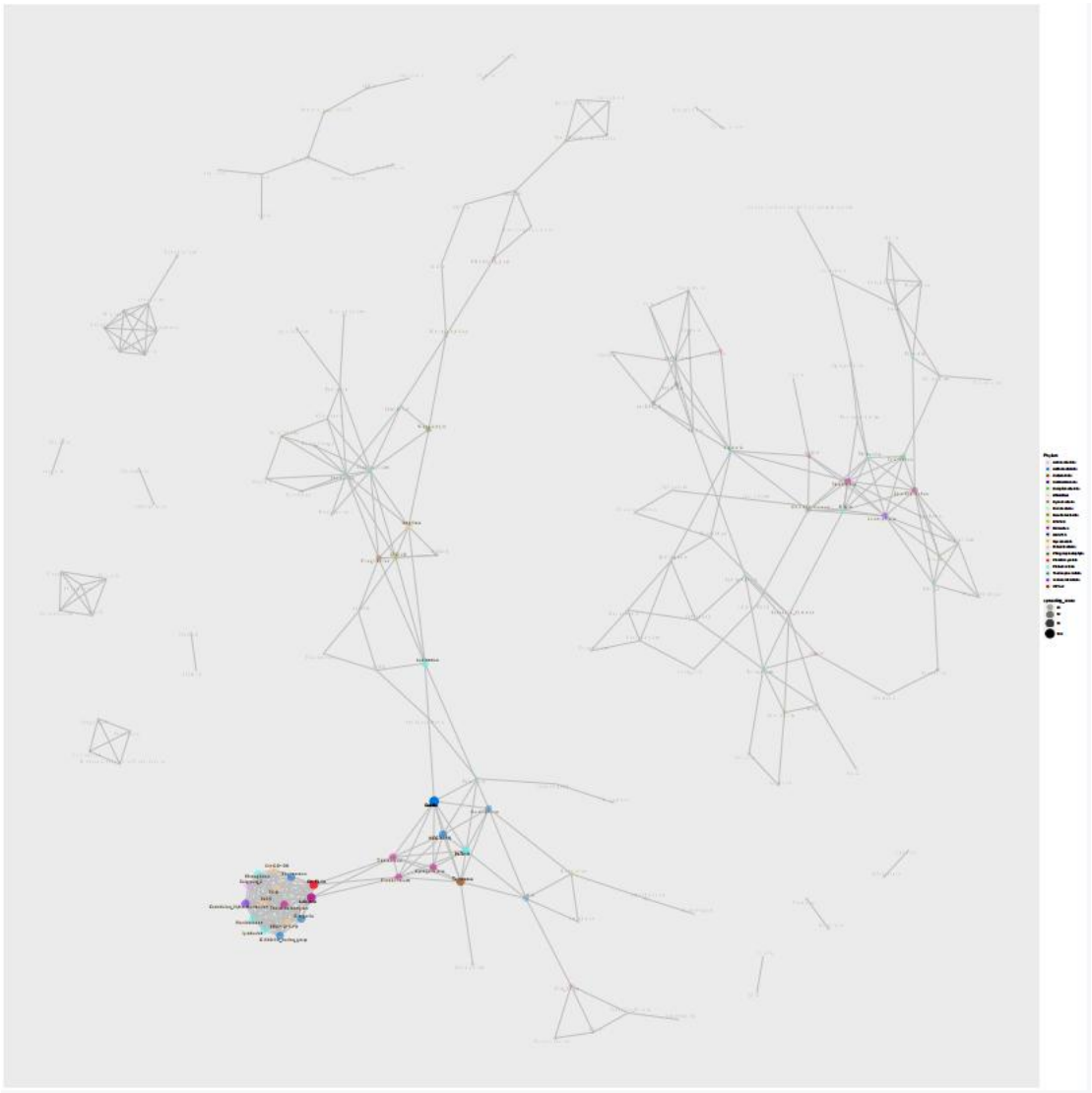
sequences and whole-genome assemblies. International journal of systematic and evolutionary microbiology, 67(5), 1613.

Appendix

Appendix I

Network images of CH sample points based on hubness and spreading values





Appendix II

Network images of COV sample points based on hubness and spreading values