UNIVERSITY OF OKLAHOMA

GRADUATE COLLEGE

STUDYING MICROBIAL COMMUNITY DIVERSITIES BY DEVELOPING HIGH-THROUGHPUT EXPERIMENTAL TECHNIQUES AND COMPUTATIONAL TOOLS

A DISSERTATION

SUBMITTED TO THE GRADUATE FACULTY

in partial fulfillment of the requirements for the

Degree of

DOCTOR OF PHILOSOPHY

By

YUJIA QIN
Norman, Oklahoma
2018

STUDYING MICROBIAL COMMUNITY DIVERSITIES USING HIGH-THROUGHPUT TECHNIQUES AND COMPUTATIONAL TOOLS

A DISSERTATION APPROVED FOR THE
DEPARTMENT OF MICROBIOLOGY AND PLANT BIOLOGY

BY

Dr. Jizhong Zhou, Chair

Dr. Meijun Zhu

Dr. Michael J. McInerney

Dr. Henry Neeman

Dr. Bradley S. Stevenson

Acknowledgements

At the end of the journey pursuing my doctoral degree in Microbiology at University of Oklahoma, I am deeply grateful for many people who have helped and accompanied me during these years, without whom, I would never get this far. First, my advisor, Jizhong Zhou, who is an intelligent and incredibly hard-working scientist. Besides the financial support over the years, Dr. Zhou has provided me guidance, encouragement, and much patience. He is the person who led me into the field of environmental microbiology, and I really appreciated the vast training I received from the projects I've got involved in and the freedom he provided while exploring the field of interest. Dr. Ye Deng, who played a role as my second advisor, is also the person I would like to thank the most. Ye provide many helps and supports during his year working with me. I learned a lot from him, such as the data analysis methods and scientific writing skills, which would also benefit me for life. I owe great thanks to my committee members: Dr. Meijun Zhu, Dr. Henry Neeman, Dr. Michael McInerney, and Dr. Bradley Steven for serving as my committee members for this long time. As a graduate student trying to develop her own background in a multidisciplinary area, I deeply appreciated their guidance throughout this whole degree period, by providing valuable suggestions for every step in pursuing this degree. Many thanks to the people for their help during my research, especially Dr. Liyou Wu, Dr. Joy Van Nostrand, Dr. Daliang Ning, Dr. Qichao Tu, who provided their kind help and cooperation in the projects I've been working on. And Missy Lee who served as a secretary in the lab and makes everything else smooth and comfortable. My final sincerest gratitude belongs to my family, for their continuous encouragements and support in my whole life. It is them who helped me to survive all the

stress for years and not letting me give up. My friends, roommates, a journey without them would have been not so worth memorizing.

Table of Contents

List of Tables

List of Supplementary Tables

List of Figures

List of Supplementary Figures

Abstract

Since the advent of high-throughput technologies, the understanding of microbial biodiversity has rapidly transformed. Amplicon sequencing of phylogenetic makers, especially 16S rRNA genes has now become a well-adopted tool to discover microbial taxonomic diversities in virtually all habitats, aquatic, terrestrial, local or global ecosystems. Although high-throughput sequencing, such as Illumina-based technologies (e.g. MiSeq), has revolutionized microbial ecology, the adoption of amplicon sequencing for environmental microbial community analysis is challenging due to the problem of low base diversity of the target region. In this study, a new phasing amplicon sequencing approach (PAS) was developed by shifting sequencing phases among different community samples from both directions via adding various numbers of bases (0–7) as spacers to both forward and reverse primers. Our results first indicated that the PAS method substantially ameliorated the problem of unbalanced base composition. Second, the PAS method substantially improved the sequence read base quality (an average of 10 % higher of bases above Q30). Third, the PAS method effectively increased raw sequence throughput (~15 % more raw reads). In addition, the PAS method significantly increased effective reads (9–47 %) and the effective read sequence length (16–96 more bases) after quality trim at Q30 with window 5. In addition, the PAS method reduced half of the sequencing errors (0.54–1.1 % less). Finally, two-step PCR amplification of the PAS method effectively ameliorated the amplification biases introduced by the long-barcoded PCR primers. The developed

strategy is robust for 16S rRNA gene amplicon sequencing, and a similar strategy could also be used for sequencing other genes important to ecosystem functional processes.

To facilitate the analysis of the data produced from the amplicon sequencing technologies, a data analysis pipeline is developed and is running to serve more than 200 users with the data processing and preliminary analysis for the amplicon sequences. The publicly available pipelines, such as QIIME(Caporaso, Kuczynski et al. 2010, Caporaso, Lauber et al. 2012) and MOTHUR (Schloss, Westcott et al. 2009), are mostly standalone services and need minimum program skills to perform the analysis. Our pipeline provides a more user-friendly interface through webpage and users will only need to click buttons rather than type command lines to perform the basic data analysis. Besides the convenient operations, the Galaxy platform provides an organized way to upload, store, track and share the data histories from different projects. The pipeline is also flexible to add new programs that are developed by others and the data source is not limited to 16S rRNAs but also functional gene amplicon sequences. The pipeline has served the research community for several years, and more than a dozen papers are published using this pipeline.

A practical application of amplicon sequencing was followed to discover the biodiversity of microbial fungal communities in six North American forests soils. The biodiversity of fungi has been studied across many habitats, but the spatial patterns of fungi diversity and the possible mechanisms behind them still need exploration. In this study, the soil fungal samples were collected from six forest sites across a wide range of latitudes in North America with a nested design in each site to uncover the diversity pattern of the soil fungal communities in forest systems. The richness of fungi follows a

clear latitudinal gradient, where temperature, precipitation, pH and nitrogen concentration also contribute to the prediction of the richness of the soil fungal communities. The compositions of fungal communities are distinct from each other across six forest sites. The main drivers of alpha diversity of fungi in forest soil are latitude, along with the mean annual temperature, precipitation, soil pH, soil total carbon, and soil total nitrogen. These seven variables can be used to predict the α-diversity of the soil fungal communities, and more than 70% variance can be explained by these variables only. As for the β-diversity, the dissimilarities among the fungal communities increases significantly as the distance between the sampling sites become larger. The distance-decay curve explains this pattern and indicates that the turnover rates of the fungal species are different in the local and continental scales. We further proved that, the key drivers of the difference in fungal community composition highly depends on the spatial scale, and the geographic distance is the major contributor to explain these differences. In summary, this study of the fungal communities in the North American forest soils has shown several patterns along with the possible drivers behind them, which presents insights into the nature of soil fungal communities.

When the advanced high-throughput technologies have enabled researchers to gain unprecedented insights of the diversity of microbial communities without culturing and identify individuals, the merely knowing the answer to “who is there” is no longer enough, the question now is to link the ‘measurable’ community structures to the ecosystem functioning. If this connection can be set up, then it is possible to understand that how the disturbances brought by the human activities and global climate change will change the ecosystem functioning carried out by microbial communities.

Functional diversity, which measures the range of things that organisms do in the surrounding ecosystem has shown its power in linking the microbial communities to the dynamics of ecosystems. In the final part of this study, we provide a framework using Rao's entropy to quantify microbial functional diversity based on GeoChip (a high-throughput functional gene array), and the phylogenetic distances between each probe is considered in the calculation. This index falls into the category of trait-based functional diversity, with the advantages of pre-selected key functional traits related to functional ecosystem designed in GeoChip. This functional diversity index can be partitioned into α- and β- diversity, which extends the understanding of functional diversity pattern into different temporal or spatial scales. The functional redundancy can also be defined following the definition of the functional diversity, which is more like a measure of gene similarity or convergence, rather than the traditionally defined 'functional redundancy' for multiple functionalities in an ecosystem. Given the hypothesis that sequence similarity leads to function similarity, the new definition of functional redundancy can reveal the redundant level of functional traits in the same gene. We applied this functional diversity framework to study the dynamic changes over a 9-month period of microbial communities in a contaminated groundwater system (with U(VI), SO_4^{2-}, NO_3^-, etc.,) after a one-time EVO (emulsified vegetable oil) amendment, which has been proven that it can effectively reduce U(VI) for a considerable time period (around one year). Using the acetate production as the measurement of EVO degradation process, the functional diversity of the key gene responsible for degradation of EVO significantly correlate with the function itself (R^2 = 0.685, p-0.021), where the other functional indices such as the gene richness did not show such a strong

relationship. When using functional diversity to profile the whole community functional structure, statistical tests also proved that the change of environmental variables does shift the community functional structure, while this connection is not as clear if using other indices to represent the community functional structures. In summary, the new framework of function diversity integrates both functional traits and their phylogenetic signals, which has been proven to be a more sensitive indicator of ecological functions than traditionally used gene richness.

Chapter 1: Introduction

1.1 Microbial biodiversity and current challenges

Biodiversity, the variability among living organisms from all ecosystems including terrestrial, marine, atmosphere and others, forms the foundation of the ecosystem services which is closely related to the wellbeing of our planet (McCann 2000, Tilman, Reich et al. 2006, Wagg, Bender et al. 2014). This complex and dynamic variation has experienced dramatically changes at the hands of humans (Vorosmarty, McIntyre et al. 2010, Bennett, Cramer et al. 2015). The change of biodiversity will have great effect upon our ecosystems (Sala, Chapin et al. 2000) and will compromise the stability and well-functioning of the ecosystems in local or even global scales (Bracken, Friberg et al. 2008, Wagg, Bender et al. 2014). Therefore, studying biodiversity can not only help further describe the image of our world with the cognition of the living organisms who shared the same environment with us, but also will provide insights to protect our living environment and keep it well-functioning against the changes of atmospheric carbon dioxide, climate, vegetation, and human activities, such as land use. Microorganisms, as the most abundant and diverse members on the planet, they contribute greatly towards the function in our ecosystems, such as global carbon cycling, nutrient availability, human health and disease development for all living organisms (Colwell 1997, Fitter, Gilligan et al. 2005, Nielsen, Ayres et al. 2011, Singh 2015, Delgado-Baquerizo, Maestre et al. 2016). Their small body size, short generation time and genetic plasticity give microorganism the capability of rapidly adaptation to the change of the environment. Therefore, the diversity of microorganisms are good bioindicators of the

perturbations of environment and ecosystems stability (Bouchez, Blieux et al. 2016, Karimi, Maron et al. 2017).

There are several ways to classify biodiversity. From the spatial scale at which the samples are taken to measure the diversity, it can be separated into alpha, beta and gamma diversity. These concepts were first introduced in 1960 (Whittaker 1960). Alpha diversity is the diversity of a relative small area at local scale, such as plot, study site, which is frequently expressed as species richness or other low-order Hill number (Tuomisto 2010). Gamma diversity is the total species diversity of a relatively large area comparing to alpha diversity, such as a landscape. Gemma diversity usually corresponds to the regional or global scale (Whittaker 1960). Beta diversity is defined as the difference or ratio between the reginal and local species diversity, that is the beta diversity can be calculated from gamma and alpha diversity. Beta diversity represents the differentiation among habitats, so it often can be represented as the pairwise dissimilarities among habitats.

In terms of species assembly, taxonomic diversity (TD) is the most commonly used diversity measurement, which plotted as taxonomic richness in species level often with some reference to temporal and geographic scale. TD describes the existence and abundance of taxonomic units, without the any consideration of the relationships between these units and related functionalities they may possess, which are often the potential goal to study biodiversity in most areas. Phylogenic diversity (PD) take the phylogenic information among the species into account. To be more specific, if the community comprised by species that are phylogenetically close to each other, the PD of this community is lower than the community consists of divergent species from the

evolutionary perspective. Functional diversity measures the functional divergence of a community using functional traits processed by the members from this community. Each functional trait can be considered as a function that the members contribute to the ecosystem.

However, the study of microorganism in the environment has not been easy due to their extremely small size, enormous diversity and complex interactions with the environment surround them. The major challenges before the high-throughput metagenomics technology was developed, is the unculturable nature of the majority microorganisms in the world. The majority of microbial diversity cannot even be detected using traditional lab techniques, so the study of the diversity pattern of microbes are restricted to extremely small scale. This kind of biodiversity studies did provide insights into some simple principles that may or may not exist when the scope is larger, but they certainly uncapable to uncover the diversity patterns of the most abundant and various organisms on earth.

1.2 Amplicon sequencing technology and taxonomic diversities

High throughput sequencing technology was inspired by the completion of human genome project in 2003, and its advent has opened a new era for the field of microbiology ecology during the last decade. The high-throughput feature of this technology provides a way to explore complex communities in depth. The high-throughput sequencing technologies are also called the next-generation sequencing (comparing to Sanger method) where sequencing reactions are produced in parallel and output enormous number of sequencing reads directly. The next generation sequencing technology (NGS) have been evolving since its birth, leading to higher and higher yield,

dropping cost, improving sequencing quality and longer read lengths (Goodwin, McPherson et al. 2016).

Amplicon sequencing, as one of the most important application of NGS, is widely applied to study the microbial composition patterns in our bodies (Grice, Kong et al. 2009), in the oceans (Sogin, Morrison et al. 2006), and in our planet (Lauber, Hamady et al. 2009). Amplicon sequencing targets for specific gene markers, such as small subunit rRNA gene (16S rRNA gene, 18s rRNA genes) and use them to profile the taxonomic structure of microbial community, because the ubiquity and conservation of these markers. High-throughput sequence reads combining with barcode indexing, have allowed investigating a large number of microbial communities in depth simultaneously, which largely expend the biodiversity study scales in the field of microbial ecology (Herlemann, Labrenz et al. 2011).

With millions of reads in a single run, computational tools to store, integrate, preprocess, and analyze these sequencing data are required extensively. There are many amplicon sequencing data analysis pipelines available for public to use, such as the popular QIIME (Caporaso, Kuczynski et al. 2010) and Mothur (Schloss, Westcott et al. 2009). However, to use such pipelines still requires users to install the tools and do a minimum command typing, and for each step, it is difficult and tedious to keep track of the parameter used and the corresponding output files without any standardization. Besides, the scale and collaborations of the metagenome research projects have become larger and more complicated, the need to share and interactively deal with the data has become a management issue if users only install the pipelines on their own computers.

Therefore, more convenient and up-to-date bioinformatic tools are always in great needs.

1.3 Functional diversities and functional gene array (GeoChip)

Functional diversity studies have received noticeably increased attentions during recent decades. Linking the biodiversity to the ecosystem processes has always been a crucial step for ecologists to understand ecosystem functioning and predict the possible effect from the loss of biodiversity caused by human activities and global climate change. Functional diversity is defined as "the value and the range of those species and organismal traits that influence ecosystem functioning" (Laureto, Cianciaruso et al. 2015). Functional diversity is based on the functional traits, which directly influence organism performance or fitness (Mouillot, Graham et al. 2013). The selection of functional traits determines what functions to be studied and how accurate can the functional diversity index explains the functional space of the species. With the definition of functional index and careful selection of functional traits, the researchers can investigate two major relations: how the species affect ecosystem functioning and adapt to the change of the environment in return (Gagic, Bartomeus et al. 2015).

1.3.1 Measuring functional diversity

In the early development of functional diversity, researchers started to focus on different ways that organisms use resources and classified species with similar patterns together assuming they would respond to environmental change similarly, and these classifications were termed "guild" or "functional groups" (Blondel 2003). At this time, the classification always relies on expert opinions, which makes the process subjective and artificial. As the global effects of human activities, climate changes became

growing concerns, interests of these "functional groups" have increased gradually. Understanding how the species react to these changes will affect ecosystem functions, instead of just studying the distribution pattern of organisms, has become research focus and thus stimulate the application of the concept: functional diversity (Laureto, Cianciaruso et al. 2015). Around 2000s, to make the comparison across studies possible, researchers introduced the idea of functional traits, so that same traits can be used and measured across different studies (Cornelissen, Lavorel et al. 2003). From then on, trait-based studies have become a popular tool for understanding the importance of the functional diversity in maintaining ecosystem function, and the response of species when the environmental changes in return (Hooper, Chapin et al. 2005, Balvanera, Pfisterer et al. 2006, Martiny, Jones et al. 2015, Perronne, Munoz et al. 2017, Colin, Villeger et al. 2018).

To estimate functional diversity, the first key step is to select appropriate traits. However, choosing traits can represent true functions and feasible to measure at the same time is not a simple task. Based on different research questions and function of interest, functional traits can be adopted, modified, and created. And the more specific and explicit the function of interest is defined, it is more likely to make reasonable and informative choices. The number of traits that used to measure functional diversity is also an important choice. If the number is too small, which means the species will occupy only a small proportion of the functional space, can lead to insufficient presentation of the function and increase the functional redundancy since species will more similar based on only a few traits. When a greater number of traits are included, the species will become more unique to each other in terms of function capabilities

(Petchey and Gaston 2006). When the variance explained by the traits selected are not increased after adding new traits, then the number of traits selected probably can make good estimation of the functions.

There are variety of approaches available in the literature to calculate functional diversity given a set of selected functional traits (Petchey and Gaston 2006, Mouchet, Villeger et al. 2010, Schmera, Heino et al. 2017). These methodology concepts using multiple traits include but not limited to: functional group richness, functional attribute diversity (FAD) (Walker, Kinzig et al. 1999), average functional attribute diversity (AFAD) (Heemsbergen, Berg et al. 2004), modified functional attribute diversity (MFAD) (Schmera, Eros et al. 2009), functional diversity (FD) (Petchey and Gaston 2002), generalized functional diversity (GFD) (Mouchet, Guilhaumon et al. 2008), functional richness (FRic) (Cornwell, Schwilk et al. 2006), Rao's quadratic entropy (Q), (Rao 1982, Botta-Dukat 2005), functional divergence (FDiv) (Villeger, Mason et al. 2008), functional evenness (FEve) (Villeger, Mason et al. 2008). Using artificial dataset representing different community assembly rules, the relationship of these indices have been proved to measure different faucet of the functional diversity, where some of the indices are highly similar (Mouchet, Villeger et al. 2010). Among these indices, Q, FDiv, FEve take the abundance of trait values into account, while others only consider the absence/presence of the trait.

Rao's quadratic entropy, defined in (Rao 1982), incorporates the pairwise distance among taxa and weighted by the relative abundance of these taxa, which takes into account the differences between traits and also the abundance difference for different traits. It has become the most frequently used measure of functional diversity,

since it not only fulfills a priori criteria (Mason, MacGillivray et al. 2003, Botta-Dukat 2005) for diversity indices, but it also quantifies the divergence and richness aspect of functional diversity (Mouchet, Villeger et al. 2010). In addition, Rao's entropy, like classical diversity indices, can be partitioned into α and β components (Ricotta 2005, Hardy and Senterre 2007, Villeger and Mouillot 2008). The partitioning process can help to reveal the function diversity patterns among and within community, and to investigate community assembly rules which may differ at different spatial levels. Traditionally, for additive petitioning, the total functional diversity (γ) is the sum of the average within-community diversity ($\overline{\alpha}$) and the among-community diversity (β). However, when decoupling the functional diversity into α and β components using Rao's entropy, the simple average of within-community diversity ($\overline{\alpha}$) might exceed the total diversity (γ), so weighted average within-community should be used to avoid negative among-community diversity (β) (Villeger and Mouillot 2008). Another important aspect of trait composition is functional redundancy (Laureto, Cianciaruso et al. 2015), which can be observed when functional diversity more rapidly reaches saturation than species richness. Functional redundancy represents the functional similarity among species, and highly similar species are usually expected to participate similar functionality in the ecosystem, which can be considered functional redundant (de Bello, Leps et al. 2007). Functional redundancy can influence the stability and resilience of a community by maintaining ecosystem functioning when loss of species diversity (Naeem 1998, Pillar, Blanco et al. 2013). Using Rao's entropy, the functional redundancy can be defined as the difference between Gini-Simpson diversity index and the functional diversity (Rao's entropy), where the former didn't consider the

dissimilarities between traits (taxa). So, when the dissimilarity between different taxa is higher, the difference between Gini-Simpson diversity and Rao's entropy be smaller, which means the functional redundancy is lower, since the taxa are distinct from each other and cannot be considered the same or redundant (Ricotta, de Bello et al. 2016).

1.3.2 Microbial functional diversity and GeoChip

Microorganisms as a most abundant and diverse members on earth, provide essential services to the ecosystem which can directly affect the wellbeing of all other living forms. Though the diversity of microorganisms is high, only very small percentage of them is recognized due to their invisible and unculturable nature, leaving a big gap in knowledge. Fortunately, the continuous development and improvement of technology, make the detecting and studying indivisible organisms possible, and start a new era of the microbial biodiversity research. Besides the enormous taxonomic diversity of microorganisms, the linkage between microbial diversity and ecosystem functions have received more and more attentions by the concern that losing microbial diversity will undermine ecosystem functions due to human activities and recent climate change. In the early days, traditional ways to study functional diversity of microorganisms are measuring certain microbial functions from various microbial communities under different conditions. Using such methods, functional diversity can be represented by, for example, microbial biomass, key enzyme activities involving nutrient cycling (Kandeler, Kampichler et al. 1996), different substrate unitization pattern using commercially viable Biolog plates (Zak, Willig et al. 1994, Preston-Mafham, Boddy et al. 2002). These methodologies are limited in the number of functions can be studies, and the activities or functions measured in situ instead of real environment are not

necessarily reflect the real microbial functions in the ecosystem. With the advent of high-throughput technologies such as microarray and sequencing, microorganisms that cannot be seen or cultivated can be detect and identified by their genetic signature, which offers great opportunities to examine the relationship between the microbial communities and ecosystem functioning. So now, the microbial functional diversity indicates the potential ability of microorganisms to express functions in the environment, which is a promising indicator of the actual microbial metabolic activities, that is the function of the system.

GeoChip is a functional gene array designed with probes targeting key genes involved in various ecosystem process (He, Deng et al. 2010, Tu, Yu et al. 2014). The newest version of GeoChip (version 5.0) contains about 1.6 million probes targeting more than 1,590 genes (Zhou, He et al. 2015) that are categorized by their roles in the ecosystem functioning, such as carbon, nitrogen sulfur and phosphorus cycling, energy metabolism, antibiotic resistance, metal homeostasis and resistance, secondary metabolism, organic remediation, stress responses, bacteriophages and virulence. Geochip is a powerful to study the function composition and structure of microbial communities, with close-format design avoiding reproducibility issue that may cause by inadequate random sampling effort (Zhou, He et al. 2015). In general, community DNAs are extracted, labeled with fluorescent dyes and hybridized with GeoChip slides. The resulting digital images are processed and translated into signal intensity for each probe, with higher signal intensity indicating higher abundance of the gene this probe targeting at. Given functional profile resulting from GeoChip analysis, gene abundance and richness (probe numbers) can be derived easily. If we each gene represents one

microbial function, then we can also calculate functional diversity for this function by analyzing the probes belong to this gene. These probes can be treated as taxa, or traits for this specific gene, and the definition of functional diversity using Rao's entropy, if the distance among these probes can be provided, then we can calculate the function diversity for this function. Using this approach, one can not only observe the functional potential using gene abundance and richness as traditional ways, but also investigate how the microbial functional composition and structure changed under different environmental conditions for each individual gene. The combination of GeoChip and functional diversity will provide a novel insight to the underlying mechanisms of the linkage between microbial communities and ecosystem functioning.

1.4 Foci of this study

As the advanced high-throughput technologies enable researchers to gain unprecedented insights of microbial communities without culturing and identify individuals, the study of microbial diversities has entered a new era. Along with the development of the field, technologies have been kept improving to produce more data with higher efficiency and accuracy; pipelines were developed to standardized the raw data processing steps and generate biological manful results that can be further interpreted and explained; new data analysis methods are invented to investigate the data in different perspectives and mining for patterns hidden behind the massive information. This dissertation aimed to contribute to the field of microbial biodiversity study in terms of technology, data analyzing pipeline and research methods.

In Chapter 2, a new phasing amplicon sequencing approach is proposed to improve the low diversity issue that causing sequencing problems using Illumnia

sequencing platform. A spacer with random length (0-7 base) is added to the primer (both forward and reversed), which will shift the sequencing phases differently to avoid the low diversity caused by the conserved region in the targeted genes. This method has been proved to successfully solve the low-diversity issue of targeted gene sequences and dramatically increase the accuracy of the amplicon sequencing results. A data analysis pipeline was development to process the amplicon sequencing data and provide preliminary interpretation of the data. The pipeline was built on Galaxy planform run by a Linux server, and it provide an interactive webpage service for users to upload, analyze, store, share, and track sequencing datasets and their analytical results. The pipeline is flexible for adding new analytical tools and is not limited to the analysis of 16S rRNA gene sequences. Mock communities from 33 know strains were used to evaluate the new phasing approach, and the data processing methods. The sequencing error rate, chimera rate, were decreased by phasing technology and the quality of the sequencing results is also improved, in terms of effective read length and number.

In chapter 3, the amplicon sequencing technology is used to study the biodiversity pattern of soil fungal communities from six forest sites in North America. ITS (nuclear ribosomal internal transcribed spacer) region is used as the phylogenetic marker and the sequence reads are analyzed through the pipeline constructed in Chapter 2. The results showed that the biodiversity of microbial fungal communities follows some basic rules that has been discovered in macroecology. One is that the α-diversity follows the latitude gradient, in other words, the soil fungal community has higher diversity when closer to equator. Another is distance-decay pattern, which indicates that as the geographic distance become larger, the soil fungal community become less

similar. We also investigated the potential mechanisms behind the α- and β-diversity patterns shown in the fungal communities. The key driver of the α-diversity, consist with the latitude gradient pattern, is the latitude, followed by temperature, precipitation, soil pH, total carbon and total nitrogen in soil. The plant richness is the most correlated factor with the fungal richness, however, it can be completely expressed as linear combinations of other environmental variables, which is why it is not contained in the final model. The drivers behind the β-diversity are different in different spatial scale. This study provides the traditional analyze methods to study microbial biodiversity and add more insights of the mechanisms behind fungal biodiversity patterns to the whole picture.

In Chapter 4, in order to link microbial community to ecosystem functioning, another aspect of biodiversity, functional diversity has been studied, and a new framework has been proposed to calculate a new functional diversity index based on GeoChip data in combine with phylogenetic linkage between the individual taxon/probe. Rao's entropy was used to combine these two pieces of information and a functional diversity can be calculated for each single gene. The diversity index can be partitioned into α- and β- diversities and extend the investigation of functional diversity pattern into different spatial scales. Functional redundancy can also be defined using this framework, though it differs from the traditionally defined redundancy, it can provide information such as gene similarity, which can also be help to understand the community assembly processes. The application of this newly development method has showed a stronger relationship between gene functional index and the corresponding ecosystem function, such as biodegradation of EVO (emulsified vegetable oil). When

using functional diversity as the unit to profile the functional structure of the whole community, the new index also reveals that the environmental variables govern the shifts of microbial functional structure, while the traditionally used gene richness did not show this pattern. In summary, the new proposed function diversity index possesses a closer relationship to the ecosystem functioning, which would help to understand how the environment change will affect the microbial functional diversity and further, the ecosystem functions.

Chapter 5 summarized the work of this dissertation and indicate the significance of this study and the contribution it made to the field.

Chapter 2: Phasing amplicon sequencing on Illumina Miseq

2.1 Abstract

Although high-throughput sequencing, such as Illumina-based technologies (e.g. MiSeq), has revolutionized microbial ecology, adaptation of amplicon sequencing for environmental microbial community analysis is challenging due to the problem of low base diversity. A new phasing amplicon sequencing approach (PAS) was developed by shifting sequencing phases among different community samples from both directions via adding various numbers of bases (0–7) as spacers to both forward and reverse primers. Our results first indicated that the PAS method substantially ameliorated the problem of unbalanced base composition. Second, the PAS method substantially improved the sequence read base quality (an average of 10 % higher of bases above Q30). Third, the PAS method effectively increased raw sequence throughput (~15 % more raw reads). In addition, the PAS method significantly increased effective reads (9–47 %) and the effective read sequence length (16–96 more bases) after quality trim at Q30 with window 5. In addition, the PAS method reduced half of the sequencing errors (0.54–1.1 % less). Finally, two-step PCR amplification of the PAS method effectively ameliorated the amplification biases introduced by the long-barcoded PCR primers. Conclusion: The developed strategy is robust for 16S rRNA gene amplicon sequencing. In addition, a similar strategy could also be used for sequencing other genes important to ecosystem functional processes

Keywords: Next generation sequencing, Low diversity sample, Amplicon sequencing, Illumina Miseq, Microbial community, Phasing primer, Microbial ecology

2.2 Introduction

Microorganisms are the most abundant diverse life forms on Earth, and they are almost everywhere (Whitman, Coleman et al. 1998). Microbial activities contribute greatly to many critical ecosystem functions. But due to their vast diversity and as-yet uncultivated nature, how to detect, identify, quantify and characterize them are some of the great challenges for researchers. In the last couple of decades, the development of high-throughput sequencing technologies has provided microbiologists ways to tackle these challenges and discover the microbial world in a whole new perspective. One of the most common application in microbial ecology is sequencing amplified gene makers (also called amplicons), such as 16S ribosomal RNA gene, fungal ITS region, nifH gene (Dethlefsen, Huse et al. 2008, Nilsson, Ryberg et al. 2009, Silva, Schloter-Hai et al. 2013) to study the phylogenetic/functional diversity and structure of microbial communities (Caporaso, Lauber et al. 2012, Faith, Guruge et al. 2013, Tromas, Fortin et al. 2017). There are various next generation sequencing (NGS) technologies are available right now, and the Illumina platform (e.g., HiSeq2000, MiSeq) has become an common option due to its lower cost, rapid analysis process, and higher accuracy (Bartram, Lynch et al. 2011, Caporaso, Lauber et al. 2012, Faith, Guruge et al. 2013, Sikkema-Raddatz, Johansson et al. 2013, Tromas, Fortin et al. 2017, Gaby, Rishishwar et al. 2018). It is anticipated that the MiSeq platform in particular will be a dominant sequencing technology for microbial ecology studies due to its great flexibility, fast-turnaround time, longer sequence reads and high accuracy (Gibson, Shokralla et al. 2014, Nelson, Morrison et al. 2014, Schirmer, Ijaz et al. 2015).

In amplicon sequencing, to decrease experimental cost and maximize the capability of sequencing technology, different community samples are often sequenced together in a single Hiseq lane or Miseq run via the use of barcodes, which are added during PCR amplification (Krueger, Andrews et al. 2011). However, low sequence diversity or unbalanced base composition in template DNA sequences are inherently problematic in amplicon sequencing with Illumina sequencing technologies, because they can affect sequence output, quality, and error rate due to problems in cluster identification, focusing, phasing/pre-phasing and color matrix estimation (Krueger, Andrews et al. 2011). Innovations such as new reagents kit (Lundberg, Yourstone et al. 2013) has been proposed to mitigate the issues, but it is still challenging. Frameshifting with different length of barcodes (3–6 bases, three bases difference) (Hummelen, Fernandes et al. 2010) and short spacers (0–5 bases) (Lundberg, Yourstone et al. 2013) have been used to shift sequences in template DNA, but these shifts are inadequate, especially for the region with continuous homopolymer. For example, there are five 'GGG', three 'GGGG', and one 'GGGGG' homopolymers within the 16S rRNA gene v4 region. Recently, longer spacers (0–7 bases) were used in a dual-indexing primer design for reducing the number of barcoded primers in multiplex 16S rRNA gene amplicon sequencing and higher quality of sequence reads were reported (Kozich, Westcott et al. 2013, Fadrosh, Ma et al. 2014). This design put spacers of 0–7 bases after indices of 12 bases in both forward and reverse primers, which are positioned after the Illumina HP10 or HP11 (Illumina, San Diego, CA, USA) sequencing primers. Therefore, the sequencing for both forward and reverse reads starts at the indices of the forward and reverse primers, sacrificing a total 24 bases of the paired end reads, which

will be essential for some long amplicon sequencing if assembly of the paired end reads is desired.

Here, we developed a new 16S rRNA gene-based amplicon sequencing strategy to ameliorate the problems associated with low diversity. In our phasing primer design, spacers of 0–7 bases are arranged in a complementary fashion in the forward and reverse primers so that the total length of the spacers is 7 bases in all paired end reads. With this spacer design, the total number of added bases between the forward and reverse primers is limited to 7 bases as to maximize the useful length of each amplicon sequence and to minimize any quality bias among sequence reads resulting from using different primer combinations. The single index of 12 bases is positioned between the Illumina adapter, which is used to hybridize the template DNA to the oligo on the Miseq flow cell, and the HP11 sequencing primer in the reverse primer. The index is sequenced separately so that it does not take spaces in the paired end sequence reads. In addition, a two-steps PCR amplification procedure is used to eliminate possible bias introduced by extra components in the long phasing primers (besides the bias introduced by target gene primers). A systematic comparison was made between Miseq runs of phasing and un-phasing methods in terms of throughput, sequence length, error rates and biases. Our results indicated that this strategy substantially increases sequence output, reads number and quality, and decreases sequencing errors, and hence can serve as a robust approach for reliably sequencing amplicons of large-scale samples from various communities.

2.3 Material and Methods

2.3.1 Samples, mock community design and DNA extraction

We've sequenced samples, including soils, ground waters, sea waters, bioreactor cultures, and saliva samples, used for PAS and non-PAS comparisons were collected from various locations and experiments. A neutral black soil planted with maize collected from Hailun, China in 2011 was used to compare one- and two-step PCR. Community DNA was extracted by freeze-grinding plus sodium dodecyl sulfate (SDS) lysis as described previously (Zhou, Bruns et al. 1996). Crude DNA extracts were purified by electrophoresis on a 0.7 % low melting agarose gel, followed by phenol extraction (Xie, Wu et al. 2012). DNA quality was assessed based on the absorbance ratios of 260/280 nm and 260/230 nm using a NanoDrop ND-1000 Spectrophotometer (NanoDrop Technologies, Wilmington, DE, USA) and the DNA concentration was quantified using a PicoGreen (Life Technologies, Grand Island, NY, USA) assay with a FLUOstar Optima (BMG Labtech, Jena, Germany).

The mock community (**Table S1.** Mock bacterial community species and details

Sequence name	Taxonomy (Phyla or class)	Source	Insert length (nt)[b]
Acidobacteria	Acidobacteria	Drinking water	1359
Actinobacteria	Actinobacteria	Wastewater reactor	1392
Bacteroidetes clone 1	Bacteroidetes	Wastewater reactor	1355
Bacteroidetes clone 2	Bacteroidetes	Drinking water	1352
Caldisericum exile	OP5	DSMZ culture collection-13637	1426
Chlorobi	Chlorobi	Surface water	1374
Cyanobacteria	Cyanobacteria	Surface water	1324
Deferribacter desulfuricans	Deferribacteres	DSMZ culture collection-14783	1410
Deinococcus indicus	Deinococcus-Thermus	DSMZ culture collection-1537	1366
Desulfurispirillum alkaliphilum	Chrysiogenetes	DSMZ culture collection-1827	1375
Dictyoglomus thermophilum	Dictyoglomi	DSMZ culture collection-396	1415

Fibrobacter succinogenes S85	Fibrobacteres	Donated by Isaac Cann, University of Illinois-Urbana Champaign	1372
Gemmatimonadetes	Gemmatimonadetes	Wastewater reactor	1360
Leptotrichia hofstadii	Fusobacteria	DSMZ culture collection-21561	1367
Mycoplasma orale	Firmicutes	DSMZ culture collection-1915	1375
Nitrospira	Nitrospirae	Wastewater reactor	1376
Persephonella hydrogeniphiia H3	Aquificae	Donated by Anne Louise Reysenbach, Portland State University	1389
Planctomycetes	Planctomycetes	Wastewater reactor	1376
Protochlamydia amoebophilia	Chlamydiae	Donated by Mathias Horn, University of Vienna	1360
Spirochaetes	Spirochaetae	Surface water	1396
Sulfurihydrogenibium yellowstonense	Aquificae	Donated by Anne Louise Reysenbach, Portland State University	1378
Synergistetes	Synergistetes	Surface water	1355
Syntrophobacter fumaroxidans	Deltaproteobacteria	Donated by Syed Hashsham, Michigan State University (DSMZ# 117)	1415
Syntrophococcus sucromutans	Firmicutes	Donated by Syed Hashsham, Michigan State University (ATCC# 43584)	1380
Syntrophomonas bryantii	Firmicutes	Donated by Syed Hashsham, Michigan State University (DSMZ# 314A)	1412
Syntrophothermus lipocalidus	Firmicutes	Donated by Syed Hashsham, Michigan State University (DSMZ# 1268)	1500
Syntrophus buswellii	Deltaproteobacteria	Donated by Syed Hashsham, Michigan State University (DSMZ# 2612A)	1413
Syntrophus gentianae	Deltaproteobacteria	Donated by Syed Hashsham, Michigan State University (DMZ# 8423)	1412
Thermodesulfobacterium commune	Thermodesulfobacteria	DSMZ culture collection-2178	1422
Thermomicrobium roseum	Chloroflexi	DSMZ culture collection-5159	1371
Thermotoga neapolitana	Thermotogae.	Donated by Claire Vielle, Michigan State University	1412
Verrucomicrobia	Verrucomicrobia	Surface water	1379
Victivallis vadensis	Lentisphaerae	DSMZ culture collection-8748	1360

[a] The mock community was a gift from Dr. Lutgarde Raskin, Department of Civil and Environmental Engineering, University of Michigan, United States of America.

[b] The insertions are near full length 16S rDNA sequences.

), which contained plasmids carrying near full length 16S rRNA gene sequences of 33 bacteria from different phyla or species at 10^9 copies/μl, was a gift from Dr. Lutgarde Raskin, University of Michigan (Pinto and Raskin 2012).

2.3.2 PCR primers and amplification

The primers used for library preparation for the non-phasing sequencing runs were gifts from Dr. Rob Knight, University of Colorado, Department of Chemistry & Biochemistry, the design of which was described previously (Caporaso, Lauber et al. 2012). These primers contained the Illumina adapter, a pad and a linker of two bases and barcodes on the reverse primers. For the two-step PCR amplification, primers [515F, 5′-GTGCCAGCMGCCGCGGTAA-3′ and 806R, 5′-GGACTACHVGGGTWTCTAAT-3′] targeting the V4 region of both bacterial and archaeal 16S rDNA without added components were used in the first step to avoid extra bias introduced by spacers and other added component.

The base diversity of sequences in sample libraries affects MiSeq amplicon sequencing in both data throughput and quality. The first 11 bases are particularly critical for cluster identification (first 7 bases) and color matrix estimation (first 11 bases). To increase the base diversity in sequences of sample libraries within V4 region, phasing primers were designed and used in the second step of the two-step PCR. Spacers of different length (0–7 bases) were added between the sequencing primer and the target gene primer in each of the 8 forward and reverse primer sets. To ensure that the total length of the amplified sequences do not vary with the primer set used, the forward and reverse primers were used in a complementary fashion so that all of the extended primer sets have exactly 7 extra bases as the spacer for sequencing phase shift.

Barcodes were added to the reverse primer between the sequencing primer and the adaptor (Additional file 2: Table S2A, B; Additional file 1: Figure S3E-G). The reverse phasing primers contained (5′ to 3′) an Illumina adapter for reverse PCR (24 bases), unique barcodes (12 bases), the Illumina reverse read sequencing primer (35 bases), spacers (0–7 bases), and the target reverse primer 806R (20 bases). The forward phasing primers included (from 5′ to 3′) an Illumina adapter for forward PCR (25 bases), the Illumina forward read sequencing primer (33 bases), spacers (0–7 bases), and the target forward primer 515F (19 bases). These primers were then used in the second step PCR.

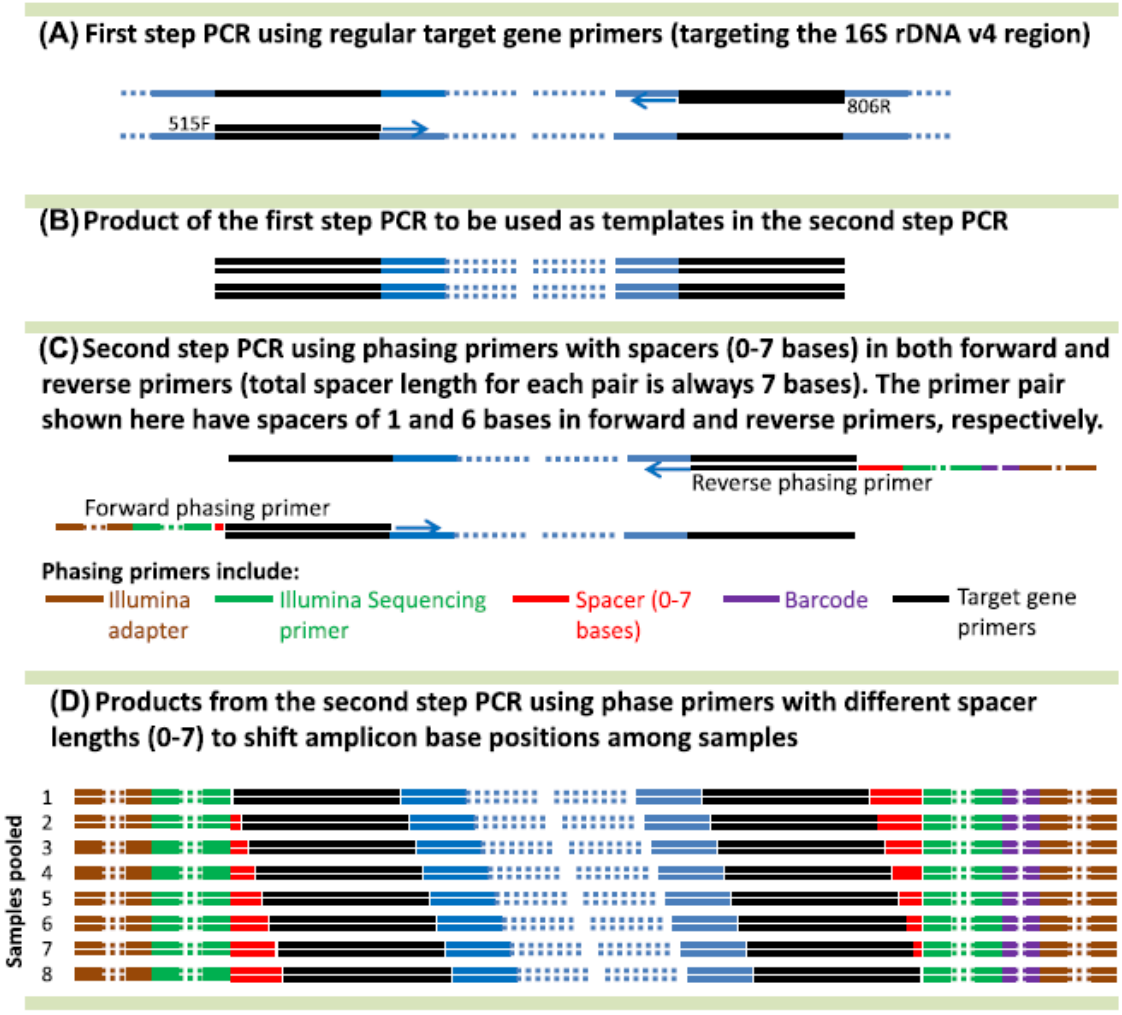


Figure 2.1 Phasing amplicon sequencing technology in four steps (a) first step PCR (b) second step PCR (c) adding spacers (d) products from the phasing technology

Tagged PCR products were generated using primer pairs with unique barcodes through either one or two-step PCR with non-phasing or phasing primers. The addition of extra components (spacers, adaptors, barcodes, etc.) to primers may introduce additional PCR bias due to their varying affinities to the upstream sequences of the target region. To minimize the potential additional bias, a two-step PCR (Fig 2.1) was used for library preparation of phasing sequencing runs. In this strategy, target-only primers were used in the first PCR reaction to amplify the target gene and that product was then used in the second PCR using primers containing all of the additional components. In the one-step PCR, reactions were carried out in a 50 μl reaction: 5 μl 10 × PCR buffer II (including dNTPs), 0.5 U high fidelity AccuPrime™ Taq DNA polymerase (Life Technologies), 0.4 μM of both forward and reverse primers, 10 ng soil DNA or 1 μl mock community of 20x dilution (start solution contained 1×10^9 copies per μl). Samples were amplified using the following program: denaturation at 94 °C for 1 min, and 30 cycles of 94 °C for 20 s, 53 °C for 25 s, and 68 °C for 45 s, with a final extension at 68 °C for 10 min.

In the two-step PCR, the first round was carried out in a 50 μl reaction as described above using target-only forward and reverse primers. Reactions were performed in triplicate and the sample amplification program described above was used except that only 10 cycles were performed. To remove residual first step PCR primers, the genomic DNA templates, and those uncompleted short PCR products, the triplicate products from the first round PCR were combined, purified with an Agencourt® AMPure XP kit (Beckman Coulter, Beverly, MA, USA), eluted in 50 μl water, and

aliquoted into three new PCR tubes (15 μL each). The second round PCR used a 25 μl reaction (2.5 μl 10 × PCR buffer II (including dNTPs), 0.25 U high fidelity AccuPrime™ Taq DNA polymerase (Life Technologies), 0.4 μM of both forward and reverse primers, 15 μl aliquot of the first-round purified PCR product). Phasing primers were used in this second round PCR with the barcode on the reverse primers. The amplifications were cycled 20 times following the above program. Positive PCR products were confirmed by agarose gel electrophoresis. PCR products from triplicate reactions were combined and quantified with PicoGreen.

PCR products from samples to be sequenced in the same MiSeq run (generally 3 × 96 = 288 samples) were pooled at equal *molality.* The pooled mixture was purified with a QIAquick Gel Extraction Kit (QIAGEN Sciences, Germantown, MD, USA) and re-quantified with PicoGreen. To keep the PCR product measurements consistent, PCR mixtures that had been previously sequenced were used as standards when a new PCR mixture was quantified. The concentration of the new PCR mixture was adjusted based on the current measurements and previous measurements of the standard PCR mixtures [adjusted new PCR mixture concentration = the measured concentration of the new PCR mixture × (the current measurement of the standard PCR mixture/the previous measurement of the standard PCR mixture)].

2.3.3 Illumina MiSeq sequencing

Sample libraries for sequencing were prepared according to the MiSeq™ Reagent Kit Preparation Guide (Illumina, San Diego, CA, USA) as described previously [5]. Briefly, first, the combined sample library was diluted to 2 nM. Then, sample denaturation was performed by mixing 10 μl of the diluted library and 10 μl of 0.2 N fresh NaOH and

incubated 5 min at room temperature. 980 μL of chilled Illumina HT1 buffer was added to the denatured DNA and mixed to make a 20 pM library. Finally, the 20pM library was further adjusted to reach the desired concentration for sequencing, for example, 625 μl of the 20 pM library was mixed with 375 μl of chilled Illumina HT1 buffer to make a 12.5 pM library. The final concentration of the library used for sequencing was determined based on the targeted cluster density. Based on manufacture protocol, the range of cluster density of 500 K/mm2–1,200 K/mm2 is recommended. The library for sequencing was mixed with a proportion of a Phix library of the same concentration. For the sequencing runs using Illumina's MiSeq Control Software version 1.1.1 and Real Time Analysis (RTA) version earlier than v1.17.28, Phix DNA spikes were adjusted to 10–20 % for phasing runs and 30–50 % for non-phasing. The incorrect hardcode matrix and phasing estimations were avoided by altering the MiSeq Configuration.xml file to use hardcode matrix and phasing/pre-phasing rates from a normal PhiX DNA run (Additional file 1: Note S1). For the sequencing runs using MiSeq Control Software v2.2.0 with RTA v1.17.28 or later, PhiX DNA was adjusted to about 10–15 % for all runs.

A 500-cycle v1 or v2 MiSeq reagent cartridge (Illumina) was thawed for 1 h in a water bath, inverted ten times to mix the thawed reagents, and stored at 4 °C for a short time until use. For non-phasing primer runs, customized sequencing primers for forward, reverse, and index reads were added to the corresponding wells on the reagent cartridge prior to being loaded as described previously [5]. Sequencing was performed for 251, 12, and 251 cycles for forward, index, and reverse reads, respectively.

Sequencing runs were monitored in real time using the Illumina Sequencing Viewer for cluster density, percentage of clusters passing filter, phasing/pre-phasing ratios, % base, error rates, % reads with quality score ≥30, and other parameters. RTA software v1.17.28 or earlier versions uses the first 4 bases for initial identification of clusters, and the first 11 bases for cluster variation. (http://supportres.illumina.com/documents/documentation/system_documentation/miseq/miseq_v2.2_software_release_notes.pdf). RTA v1.18.42 uses the first 7 bases for cluster identification and the first 11 cycles for color matrix estimation (http://supportres.illumina.com/documents/documentation/system_documentation/miseq/miseq-updater-v2-3-software-release-notes.pdf).

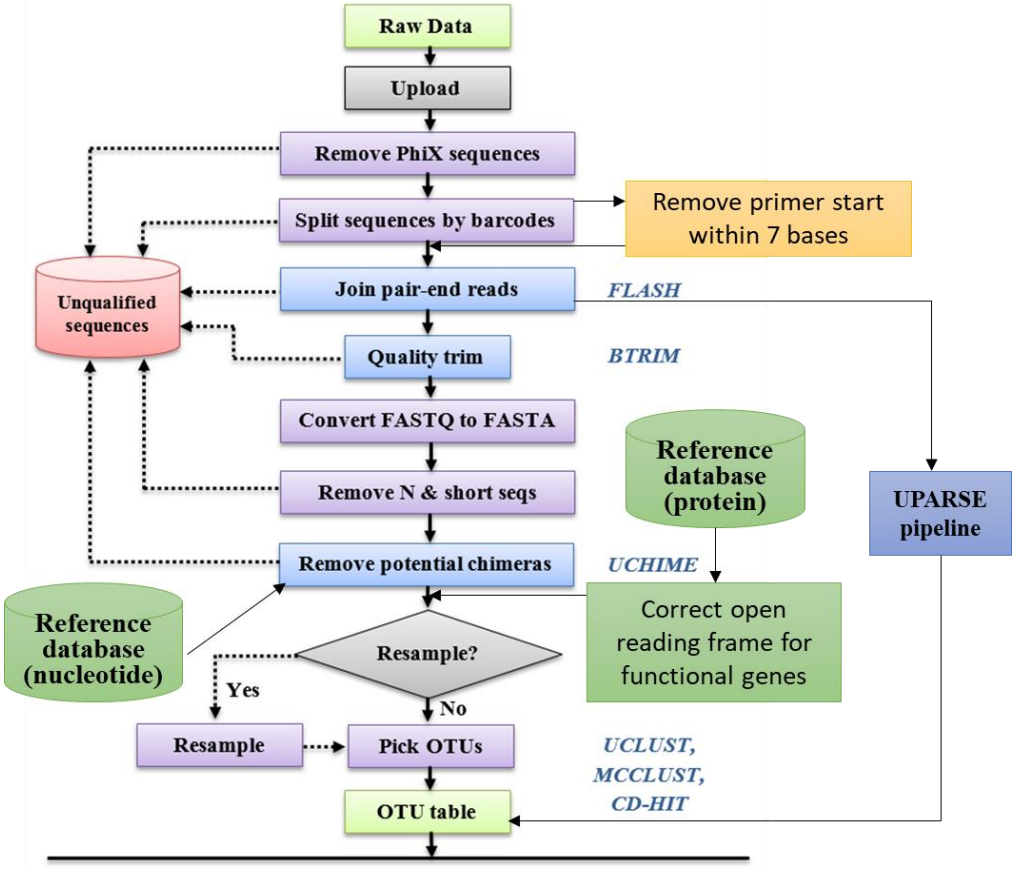
2.3.4 Data analysis and amplicon sequence data analysis pipeline

To analyze the amplicon sequencing data, a series of data processing procedure needs to performed to get meaningful biological information out of the data. There are many amplicon sequencing data analysis pipelines available for public to use, such as the popular QIIME (Caporaso, Kuczynski et al. 2010) and Mothur (Schloss, Westcott et al. 2009). However, to use such pipelines still requires users to install the tools and do a minimum command typing, and for each step, it is difficult and tedious to keep track of the parameter used and the corresponding output files without any standardization. Besides, the scale and collaborations of the metagenome research projects have become larger and more complicated, the need to share and interactively deal with the data has become a management issue if users only install the pipelines on their own computers. Therefore, we developed a data analysis pipeline for amplicon sequencing data analysis based on the Galaxy (Blankenberg, Gordon et al. 2010) platform, which allowing users

without any programming experiences to perform the analysis and manipulate the data using a point and click interface. The pipeline was installed in our Linux server with all the selected public toolkits and self-written scripts integrated into it, and the users can access the pipeline through the website http://zhoulab5.rccc.ou.edu:8080 after setting up accounts from the administers. The user can upload the raw sequencing data into the pipeline and starting to process them by selecting the tools listed aside. For each step, after setting up the input files and parameters required by the tool, the running process will automatically store the parameters in the output files, which will appear in the data history panel after the computational job is done. And once the processing steps and the parameters in each step are decided, one can create workflows to connect every step and the next time there's a similar dataset, users can run this workflow by one click and get the final results. This feature in Galaxy gives the flexibility for users who would like to explore tools and parameters to optimize the results, but most importantly, provides a convenient way for users who just want to get the final results without any complications. The user accounts information is protected by passwords but the users can share the data histories (a dataset with all the intermediate and final result files) through the pipeline by providing the other user's information (email address in our case). In this way, not only we can save the extra space to store the shared data, but also it allows users to share the whole processing steps including the data analysis tools with the parameters they use, which will avoid confusions and give the whole data processing procedure more clarity and transparency.

The major processes included in our pipeline are shown in Figure 2.2. There are two major steps in the pipeline: filtering the sequences and generate operational

taxonomic units (OTUs). In the first part, the sequences are first trimmed with their qualities, followed by checking for short and ambiguous fragments, and finally checked for chimeras. The references used for chimera checking are stored in the data libraries of the pipeline so that users can access them easily, which provides a standardization for all the datasets get processed through this pipeline. For the second part, there are many algorithms can be used to classify sequences into similar taxonomic groups, and we list several most commonly used ones: UCLUST (Edgar 2010), UPARSE (Edgar 2013), CD-HIT (Fu, Niu et al. 2012), and McClust (Scrucca, Fop et al. 2016).



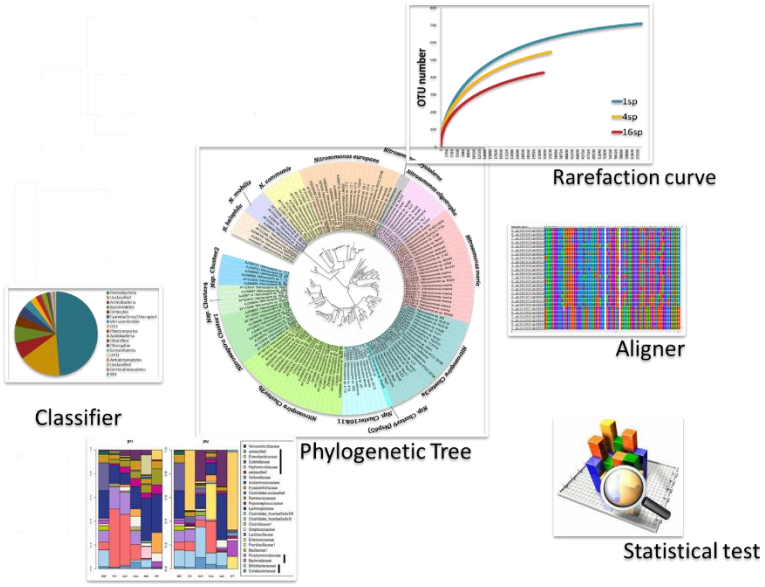


Figure 2.2 Miseq amplicon sequencing pipeline flowchart. (a) data pre-processing from reads to OTUs (b) basic statistical and phylogenetic analysis

Our pipeline also adds the ability to process functional gene amplicon sequencing data other than the traditional taxonomic markers such as 16S rRNA and ITS. To process the sequences for functional gene amplicon data, an additional step should be made during the filtering process: checking the open reading frames (Wang, Quensen et al. 2013). This will ensure the correctness when the DNA sequences are translated into protein sequences, and it can also remove some erroneous sequences, such as those contains termination codon in the middle of the sequences. And we also keep tracking of the reference sequences that can be used for different functional genes to allow accurate chimera checking steps and future classification steps.

The analysis of the OTUs are the most important part for the researchers to solve the corresponding research questions. We provide some basic tools which can be directly performed after the OTU table is generated. For example, we provided the tools that can generate the rarefaction curve based on the rarefied sequencing number and corresponding OTU numbers, which can be used to evaluate if the sequencing depth is

enough to cover the possible species in the environment and the calculated Chao value can also be as a measure of alpha diversity in each group of samples. The taxonomic classification is another necessary tool for users to assign possible taxonomic information to their sequences based on what is already known to the academic public. The RDP classifier (Wang, Garrity et al. 2007) are used for such a task for the and resampling. These tools can help users to get a general picture of the data and provide direction for more detailed and specific analysis in the future.

2.4 Results

2.4.1 Basic sequencing properties using phasing strategy

The V4 version of the 16S rRNA gene is commonly targeted for sequencing with the primer set 515F and 806R, which has high sequence coverage for both bacteria and archaea. This will produce an approximate 292 base-pair fragment including these two primers and a 253 base-pair amplicon excluding the primer sets. For optimal sequencing results using Illumina sequencing, the base diversity across a set of amplicon sequences would have an even diversity at each position so that each base (A, T, G, C) would be present in 25% of the sequences at any given position. However, the base diversity in this region of the 16S rRNA gene is very low. Of the first 100 base positions, 63% and 79% of positions in the forward and reverse sequences have one base with frequencies greater than 75% respectively. To overcome this problem of unbalanced base distribution, we use the strategy of a complimentary spacer pair containing a variable number of bases (0-7 bp, but always equaling 7 bases between the two) inserted in both the forward and reverse primers between the sequencing and target amplification sections of the primers. In this way, the sequencing phase will shift among different

community samples and thus increasing the base diversity at individual positions. After adding these spaces, the base composition in this region is more balance and the difference in nucleotide frequency for most positions is less than 30% **Error! Reference source not found.Figure 2.2** (a, b). The frequencies of the 4 bases in the first 12 bases before and after the primer shift in both forward and reverse reads are shown in the **Error! Reference source not found.** (c, d, e, f), and it is clear that this phasing amplicon sequencing strategy substantially improved the base composition balance.

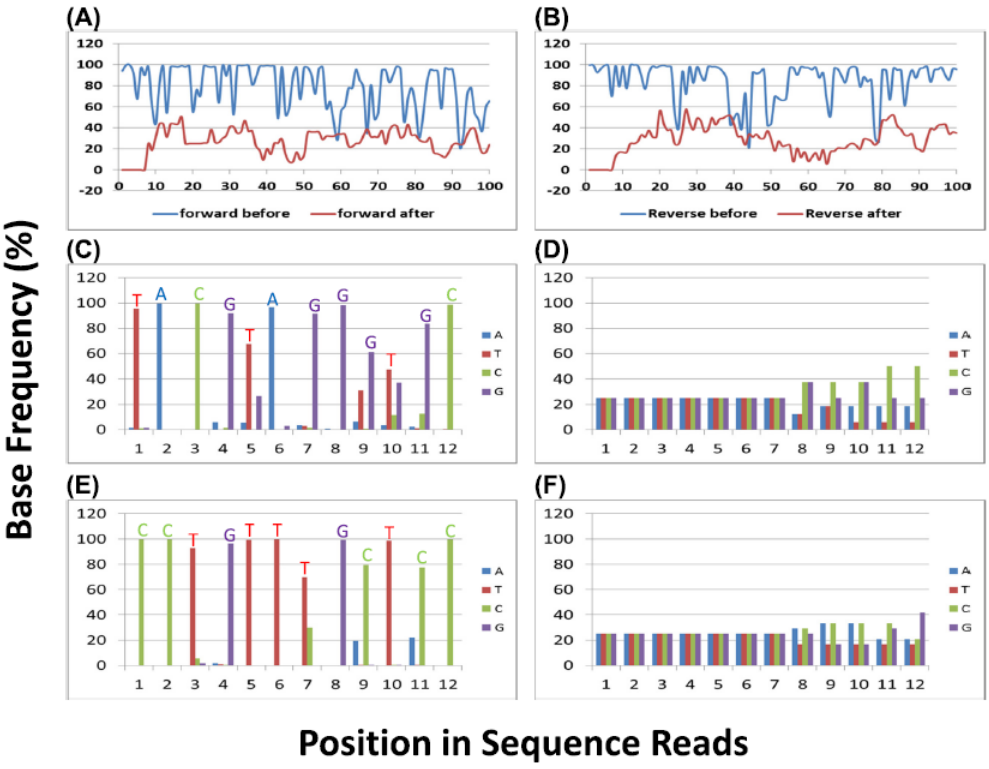


Figure 2.3 Impact of phasing primers on base frequency distributions. Differences between the maximum and minimum base frequencies at each sequence position were estimated before and after primer shift for forward (a) and reverse (b) sequences. Base frequencies of the first 12 positions of the forward sequences before (c) and after (d) primer shift, and the reverse sequences before (e) and after primer (f) shift.

2.4.2 Effective reads number and read lengths

To determine whether this PAS (phasing amplicon sequencing) strategy is consistently better than non-PAS in terms of sequence output, sequence quality and effective read length, the experimental data from different PAS runs were analyzed. These sequencing runs were used to determine the diversity of 8.731 microbial communities from diverse habitats such as soil, sediment, groundwater, bioreactors, wastewater treatment plants and human oral and guts. The percentage of sequence clusters passing the filter decreased for both PAS and non-PAS runs as the cluster density goes higher, but the PAS runs decrease slower (smaller slope) than the non-PAS runs Figure 2.4a. At the same time, the number of sequences reads also increased more when using the PAS strategy Figure 2.4a. In addition, the average percentage of bases with > Q30 at the last cycle was significantly higher ($p < 0.001$) for PAS runs (forward, 93.5 %; reverse, 88.4 %) than for non-PAS runs (forward, 86.3 %; reverse, 78.5 %) (Figure 2.4b). These results indicated that the PAS method provided high resolution for sequence cluster identification, and therefore, maximized the sequence read output, and significantly improved sequence read quality due to the balanced fluorescence signal intensity.

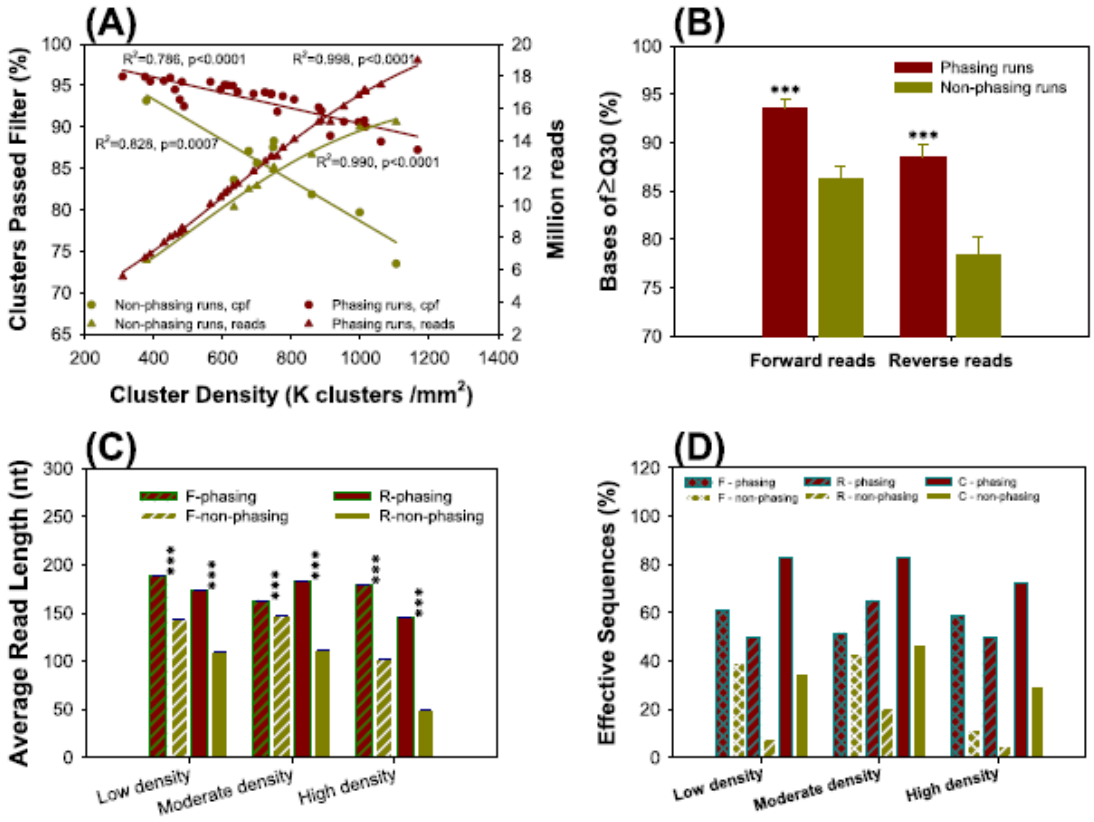


Figure 2.4 Sequence output, read length and sequence quality comparisons between the PAS and non-PAS approaches.

The PAS method was further evaluated by comparing the average read length after quality trim at Q30 and Q20 with the trimming window set at 5 or 2. The percentage of effective sequence reads, which refer to those sequences for which at least 80 % of all bases in the theoretical length have scores of >Q30 or >Q20 (e.g. 200 bp for 2 × 250 bp paired end reads), were also evaluated. The average read length for both forward and reverse sequences were significantly longer after quality trimming in PAS runs than in non-PAS runs. This was especially obvious at high cluster densities and at Q30 with the quality trimming window set at 5 (**Figure 2.4c**).

More importantly, the percentage of effective reads were considerably higher for PAS runs than for non-PAS runs for both forward and reverse sequences and for combined full-length sequences (253 bp) at all cluster densities compared, particularly at high sequence cluster densities and at Q30 for the reverse reads (**Figure 2.4d**).

2.4.3 Error rate analysis using mock communities

To determine whether PAS affects sequencing error, a mock community containing full length, plasmid-borne 16S rDNA sequences from 33 (Table S1) different bacterial phyla or classes (Ahn, Costa et al. 1996) was sequenced using both PAS and non-PAS methods (both sequencing runs were performed after the Illumina RTA software was upgraded to version 1.17.28). The relative abundance of the strains from the sequencing results and their expected value in three communities (Bm1, Bm2, Bm3) with different proportional compositions are shown in Figure 2.5. There are three strategies used: non-PAS, PAS with one-step PCR and PAS with two-step PCR. A two-step PCR amplification strategy is to amplify the target gene with standard primers at a low cycle number (e.g. 10 cycles followed by a second PCR amplification using the PCR products from the first step PCR and long barcoded primers with spacers.

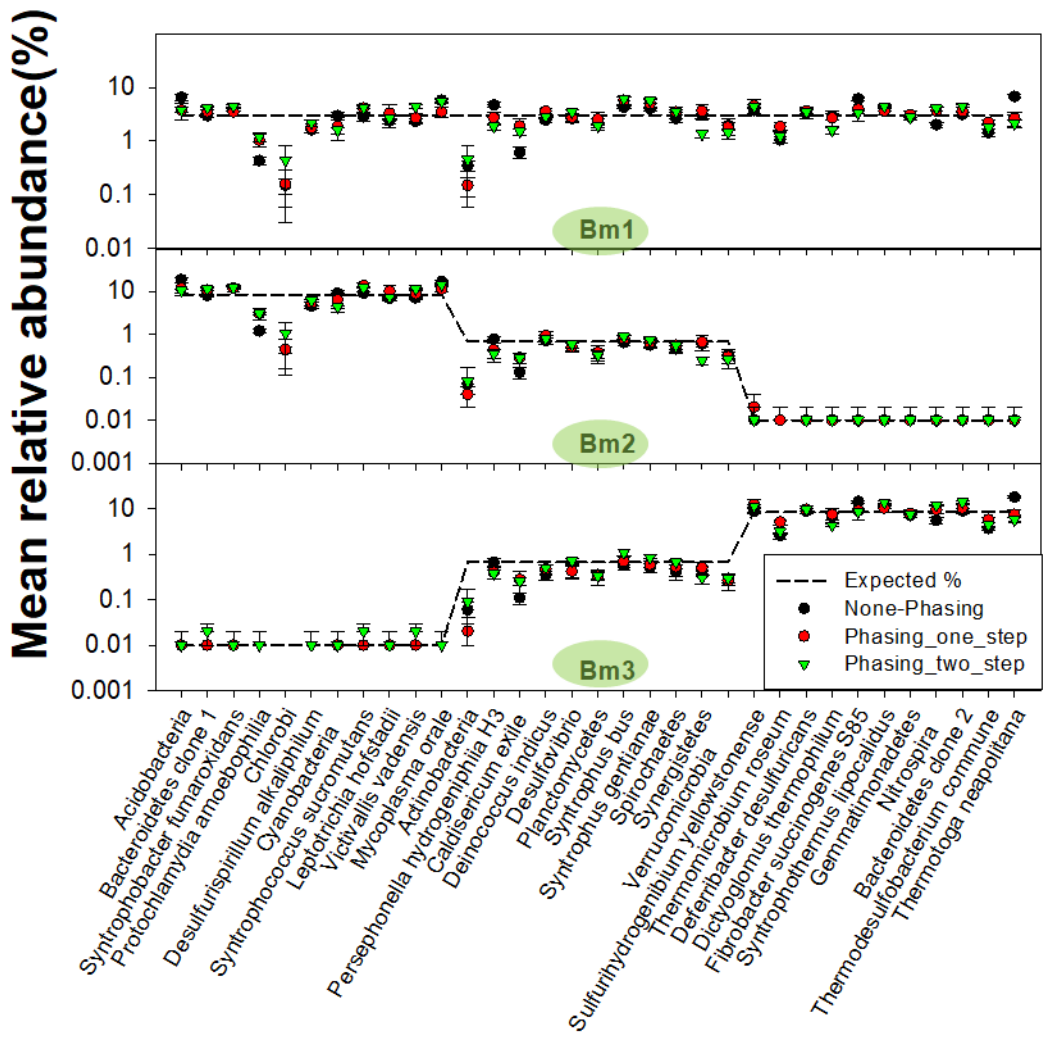


Figure 2.5 Three mock communities with different member distribution to detect PCR bias among different bacteria strains due to target gene primer preference

From the correlations between the real relative abundance and their expected values (Table 2.1), the sequencing results from the Bm1 community, which has 33 trains evenly distributed, does not have significant correlations with their expected values, whereas the other two communities have much more consistent results as expected using different sequencing strategies. And from Table 2.1, the PAS strategy using one-step PCR seems can produce community distributions closer to the real distribution than the other two strategies (non-PAS and PAS with two-step PCR). The

correlations between the three strategies (**Error! Reference source not found.**) shows that the PAS with one-step and two-step PCR are the strategies that produce the more similar results (>86.33%).

Table 2.1 Pearson correlations between mock community stain relative abundances and their expected values

Mock Community	Sequencing Strategy	r	p-value
Mock1 (Bm1)	old primer	0.0000	1
	one-step	0.0000	1
	two-step	0.0000	1
Mock2 (Bm2)	old primer	0.7799	<0.001
	one-step	0.8614	<0.001
	two-step	0.8504	<0.001
Mock3 (Bm3)	old primer	0.8909	<0.001
	one-step	0.9518	<0.001
	two-step	0.8779	<0.001

The error rate was calculated during every data processing step and based on the results, the PAS method can reduce sequencing errors. The average sequencing error rate of the raw sequence reads was significantly lower ($p < 0.0001$) for PAS than non-PAS runs (1.17 vs 1.71 % for forward sequences, 0.77 vs 1.87 % for reverse sequences). Much higher error rates were observed for non-PAS runs both before the 100th cycle and in the last 97 cycles. The higher raw sequence error rates for both forward and reverse reads in the non-PAS run was comparable to other reported error rates (Kozich, Westcott et al. 2013). Also, although sequence quality trimming significantly reduced error rates for all the approaches, error rates were still considerably higher for non-PAS than PAS runs. In addition, due to higher sequencing errors and subsequently stricter quality trimming, the percentage of effective sequence reads and combined sequences was substantially lower for non-PAS runs than PAS runs. These results indicate that the PAS method not only increased the number of effective sequence reads and read length

but also reduced sequencing errors. One way that the PAS method reduces sequence error rates could be the higher quality of the sequencing reads obtained using this method. Another reason could be that PAS has a relatively lower percentage of chimera formation during PCR amplification due to fewer amplification cycles at both amplification steps and preliminary evidence indicates that fewer chimera are present with PAS.

2.4.4 Potential bias source for OTU composition in mock communities

The final OTU number (hypothetical species) in the mock communities from all the sequencing strategies are higher than 33, which indicating there are non-expected sequences generated from the sequencing process. Since we have proved that the PAS have dramatically increase the quality of the sequences from the above sections, the major concern is what kind of bias exist in the sequences produced by PAS. The first possible step that can introduce the bias are the PCR implication process. As shown in **Figure 2.6a**, if using only one-step PCR, the community composition will be separated into 8 groups, where each group has the same spacer added (0-7 bp length). This indicates that using only one-step PCR will generate sequences depend on the spacer added, instead of the original targeting region. When using two-step PCR, the 8 groups appear no substantial differences in the final community structure, which means the second step of PCR have reduce the possibility of the artificial sequences generated from the PAS spacers.

The second and the most possible bias source the chimeras **Figure 2.6b** that also formed during the PCR process. The real chimera rate is estimated by using the UCHIME (Edgar, Haas et al. 2011) and providing the 16S rRNA sequences of the 33

strains used in our mock communities. As in the final results that used to generate OTUs, the chimeras are removed by using UCHIME and the 16S rRNA Greengene (DeSantis, Hugenholtz et al. 2006) database, to represent the normal situation that the community members are unknown in advance. For sequences before any quality trimming, the non-PAS and one-step PCR strategy produces as much as 10% chimeric sequences, while the two-step PCR strategy has a chimera rate slightly less than 6%. After applying UCHIME to remove the potential chimeras without any prior knowledge, the chimera rate will drop below 2% using the two-step PCR strategy, which the other two strategies will have around 3% chimeras in the sequences.

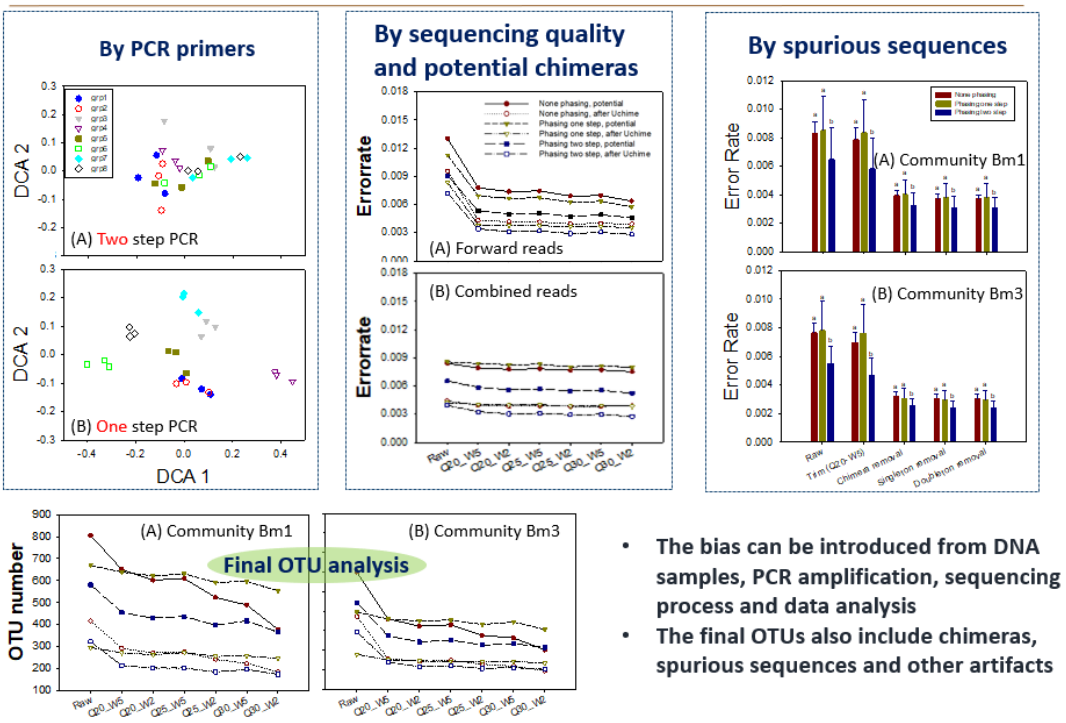


Figure 2.6 The bias introduced from various sources: DNA samples, PCR amplification, sequencing process and data analysis

The sequencing quality is another reason there are bias in the sequencing results, since sequencing error will generate sequences that different from the targets and will appear as different strains in the final OTUs. The most important reason for pre-processing the

sequences in the first place is to trim or discard any sequence reads that have low qualities, which can be estimated by the QC score from the Illumina sequencing results. These scores are stored in the FASTQ format along with the sequence content. The trimming process will remove or trim the sequence regions below certain quality threshold, and the averaging window of the quality scores is also applied to represent possible strategies to clean up low quality sequences. Normally people use Q-score of 20 and window size of 5 as the standards, which is comparatively less strict than other criteria people used in the publications. When more stringent trimming criteria is used, the error rate of the sequences should decrease, but the errors cannot be eliminated by using most stringent criteria. From the results, the criteria of Q20 and window size 5 can perform equally well as other criteria in terms of error rate in both PAS strategies, so it is not necessary to use higher standards to trim the sequences. Spurious sequences, which appears much less frequently in the samples than normal sequence, such as singletons and doubletons, can be viewed as noises as well. Even though, the reason of the generation of these spurious sequences cannot be specifically distinguished, we can remove these potential errors by removing these sequences. The error rate and OTU numbers are both decreased after removal of singletons, and since there are only a few doubletons exists in our sequences, the effect of doubletons removal is not significant.

2.5 Discussion

The phasing amplicon sequencing overcomes the low diversity and unbalanced base distribution issue, which will cause problems to form correct clusters during the Illumina sequencing processes. There are other efforts have been made to solve this

problem as well, such as by shifting the sequencing phases of amplicons by using staggered barcodes (3-6 bases) or spacers of (1-5 bases), but these methods achieved not sufficient sequence position shifts based on the simulation of the base distribution after adding bases to the 5' end of the primers. Using a 1-5 base spacer, there would be only 6 primers available (i.e. 0 bases, 1, bases, etc.), so the base distribution would still be unbalance even in the first base position, since 6 is not a multiple of 4 (the number of base available). A similar problem exists with the 3-6 staggered base barcode design. These findings suggest that a larger frame shift of at least 8 bases would be necessary to increase base diversity across the length of the entire amplicon. An additional concern is that using primers of varying length will results amplicon sequences of different length and quality bias among amplicon sequences due to their length differences. So, to address these issues, the PAS strategy developed here uses a complimentary spacer pair containing a variable number of bases (0–7 bp, but always equaling 7 bases between the two) inserted in both the forward and reverse primers between the sequencing and target amplification sections of the primer to shift sequencing phases among different community samples, increasing the base diversity at individual positions.

The difference between the number of effective combined sequences in the PAS and non-PAS methods was less than that between either the forward or reverse reads. This was most likely due to the relatively short amplicons generated from the 16S rDNA v4 region. Short reads are still a concern for amplicon sequencing with Illumina platforms even with the 2x300 bp paired end kit. If there is a relatively low base diversity, read length after quality trimming will be much shorter than expected,

especially when the quality trimming is done under highly stringent conditions, e.g. Q30. For many functional genes, such as nirK, nirS, amoA, and dsr, it is difficult to find primers to generate amplicons of appropriate length, so relatively longer amplicons (over 500 bp) must be selected. The results here indicated that PAS method effectively improved sequence read quality and length, which are critical for sequencing longer amplicons, assembling paired-end reads and increasing overall sequencing accuracy.

Since spacers and other components were added to the phasing primers before the target primer sequences, additional PCR amplification bias could be introduced. Using two-step PCR should reduce biases because the standard primers do not have added components, and when using the PCR products as target in the second run, the targets will not have up- or down-stream sequences to avoid biases introduced by the added components. The results indicated that the long primers with added components did introduce extra amplification biases with one-step PCR amplification while no apparent bias was introduced by the two-step PCR amplification. In addition, PCR amplification bias among technical replicates was also present with the one-step PCR when primers without spacers were used (data not shown). Therefore, the use of a two-step PCR approach is necessary if phasing primers or primers with added components are used for amplicon library preparation.

2.6 Conclusion

In summary, although the Illumina MiSeq and other high-throughput sequencing technologies are promising and powerful tools, adopting these technologies for analyzing microbial communities is challenging. A novel amplicon sequencing approach was developed by shifting sequencing phases among different community

samples from both directions via adding a total of 7 bases to both forward and reverse primers as spacers. Our results indicate that this approach effectively increases raw sequence throughput, read quality and effective read sequence length, and reduces sequencing errors. Analysis of MiSeq sequencing runs showed that PAS provides a robust approach for reliably analyzing microbial communities of diverse composition from a variety of habitats. In addition, our results indicate that a two-step PCR amplification strategy effectively ameliorates PCR amplification biases introduced by the use of long barcoded PCR primers. The use of a single barcode makes it easy to utilize the complementary phasing primers among samples, but multiplex amplicon sequencing requires a large number of barcoded primers, increasing the upfront costs of this method. However, despite this initial outlay, the cost per sample for the PAS method is similar to other methods. After a careful comparison of the PAS method described in this paper and other phasing methods (Hummelen, Macklaim et al. 2011, Kozich, Westcott et al. 2013, Lundberg, Yourstone et al. 2013, Fadrosh, Ma et al. 2014), the PAS method has the following unique features: i) sufficient sequence position frame shift among samples to increase base diversity across the entire sequence; ii) minimum base sacrifice by sequencing barcodes in separate reads (index reads); iii) a complementary spacer design that adds a combined 7-base spacer to both the forward and reverse primers, minimizing the total number of bases added, maximizing the amplicon sequence length, and avoiding quality biases caused by differences in amplicon sequence lengths; iv) a two-step PCR strategy that eliminates the potential extra PCR bias caused by added PCR primer components, v) lower PCR cycles in both first and second step PCR to reduce chimeras. In addition, this study is

the first time to systematically and thoroughly evaluate a phasing method for Miseq amplicon sequencing in terms of data output, sequence quality, error rate, and bias. While this strategy was developed and tested on the 16S rRNA gene, it has also been used successfully on ITS for fungi, 18S rRNA genes for protist, and other functional genes including bacterial and archaeal amoA, nifH, mcrA, and pmoA (not shown here), indicating its applicability for sequencing many different genes.

Chapter 3: The diversity pattern of soil fungal microbial community in North America forest systems

3.1 Abstract

The diversity of fungi has been studied in studied across many habitats, but the pattern of fungi diversity still needs to be revealed. In this study, the soil fungal samples were collected from six forest sites across a wide range of latitudes in North America with a nested design in each site to uncover the diversity pattern of the soil fungal communities in forest systems. The richness of fungi follows a clear latitudinal gradient, where temperature, precipitation, pH and nitrogen concentration also contribute to the prediction of the richness of the soil fungal communities. The compositions of fungal communities are distinct from each other across six forest sites. The main drivers of alpha diversity of fungi in forest soil is latitude, along with the mean annual temperature, precipitation, soil pH, soil total carbon, and soil total nitrogen. These seven variables can be used to predict the α-diversity of the soil fungal communities, and more than 70% variance can be explained by these variables only. As for the β-diversity, the dissimilarities among the fungal communities increases significantly as the distance between the sampling sites become larger. The distance-decay curve explains this pattern and indicate that the turnover rates of the fungal species are different in the local and continental scales. We further proved that, the key drivers of the difference in fungal community composition highly depends on the spatial scale, and the geographic distance is the major contributor to explain these differences. In summary, this study of the fungal communities in the North American forest soils have shown several patterns

along with the possible drivers behind them, which presents insights to the nature of soil fungal communities.

3.2 Introduction

Fungi are eukaryotic microorganisms that play fundamental ecological roles as decomposers, mutualists, or pathogens of plants and animals; they drive carbon cycling in forecast soils, mediate mineral nutrition of plants, and alleviate carbon limitations of other soil organisms. Fungi comprise some 100,000 described species, but the actual extent of global fungal diversity is estimated at 0.8 million to 5.1 million species (Fierer, Strickland et al. 2009). The biomass and relative proportions of microbial groups, including fungi, co-vary with the concentration of growth-limiting nutrients in soils and plant tissues. Such patterns suggest that the distribution of microbes reflects latitudinal variation in ecosystem nutrient dynamics. Richness of nearly all terrestrial and marine microorganisms in negatively related to increase latitude – a pattern attribute to the combined effects of climate, niche conservatism, and rates of evolutionary radiation and extinction (Hillebrand 2004). Although morphological species of unicellular microbes are usually cosmopolitan (Finlay 2002), there is growing evidence that the distribution of microorganisms is shaped by macroecological and community assembly process.

Since the high-throughput sequencing technology has enabled researchers to detect the hidden diversity of microorganisms , the fungal diversity has been studied for different taxonomic and functional groups (Nguyen, Williams et al. 2016), extreme environment (Grum-Grzhimaylo, Georgieva et al. 2016), airborne species (Woo, An et al. 2018) or human related groups (Sharpe, Bearman et al. 2015). However, the fungal diversity patterns and the possible mechanisms behind them are still need more data and evidence to uncover. A global fungal distribution survey (Tedersoo, Bahram et al. 2014)

was conducted and samples were collected from various environments at large geographic scale. The study showed that the distance from equator and mean annual precipitation were best predictor of soil fungal richness, while other environmental variables may drive the distribution of different taxonomic or functional groups. Temperature also has been shown to be a decisive factor of the fungal richness through the maritime Antarctic, the most rapidly warming region in response to the recent climate change (Newsham, Hopkins et al. 2016). Soil pH was also an important factor that can shape the fungal community along an altitudinal gradient (Wang, Zheng et al. 2015). These independent studies have revealed some part of the complexity behind the diversity pattern of fungi and there are no standard conclusions can be drawn for all the fungal communities.

In this study, we used a dataset collected from the soils of six forest sites across North America. The sampling sites were designed in a nested way so that the existence of the area-species pattern, or the distance-decay pattern can be easily detected and evaluated. The six forests are from various locations with different latitudes, average annual temperatures, annual precipitations, soil pHs, and other environmental variables. These factors can be used to build models to predict soil fungal richness while distinguish which is the dominant factor that drives the diversity pattern of fungal. The patterns of β-diversity can also be examined when the differences between the microbial fungal communities can be calculated and used as the indicator of fungal β-diversity. In summary, we are trying to uncover the diversity patterns of soil fungal community and understand what are the main factors influence these patterns.

3.3 Material and Methods

3.3.1 Six forest sites and sampling strategy

The soil samples were collected in a continental-scale survey from six forest sites in North America, as illustrated in Fig 3.1. Since the soil microbes can be locally adapted to edaphic characteristics at the scale of only a few meters (Belotte, Curien et al. 2003), and also can respond to the environmental factors as mass effects in a larger scale (Logue and Lindstrom 2010), we sampled the soils at multiple spatial scales to quantify the microbial fungal diversity in both small and large-spatial scales.

Six forest sites located in North America are: H.J. Andrews (AND, coniferous forest, 44°12′44.2″N, 122°15′19″W), Coweeta (CWT, deciduous forest, 35°3′37.2″N, 83°25′49.02″W), Harvard (HFR, deciduous forest, 42°32′16.08″N, 72°11′24″W), Luquillo (LUQ, tropical rainforest, 18°18′N, 65°48′W), Niwot Ridge (NWT, alpine tundra, 39°59′24″N, 105°22′48″W) and Barro Colorado Island (BCI, tropical rainforest, 9°09′N, 79°51′W). These selected sites represented typical forest ecosystems in North America, from boreal to tropical forests. The latitudes range from 9°N to 44°N and the temperature varies from 2.5°C to 25.7°C. The sites can also be differentiated by other variations including average annual temperature, plant species richness, annual precipitation, soil moisture, and pH, as shown in Table 3.1. The mean temperature and average annual precipitation were calculated from the hourly temperature and annual precipitation data collected through the nearest weather stations on sites.

At each forest site, 21 soil samples were collected at meter-scale using a nested design at distance of 1, 10, 50, 100, 200m in cardinal direction (Figure 3.1), where at each sampling spot, 9 soil cores were collected evenly from a 1-m^2 area (~10cm depth,

Oakfield Apparatus Company model HA). Soils were kept on ice in the field, then at −20 °C (LUQ, CWT, AND and NWT) or −80 °C (BCI and HFR) until shipped overnight on dry ice to the Institute for Environmental Genomics at the University of Oklahoma.

Table 3.1 Summary of site characteristics for the six forest sites

Sites	Latitude (°N)	Elevation (m)	Mean temperature (°C)	Precipitation (mm)	Soil moisture (%)	Soil pH	Plant species number
BCI	9.16	157	25.71	2383.0	31.43	5.87	263
LUQ	18.32	386	23.62	3069.2	40.53	5.06	93
CWT	35.05	864	12.62	1853.8	30.28	4.72	49
AND	44.23	860	8.94	1587.4	36.88	5.28	18
HFR	42.54	356	8.27	1128.7	34.35	3.84	25
NWT	40.04	3186	2.50	481.6	16.00	4.83	5

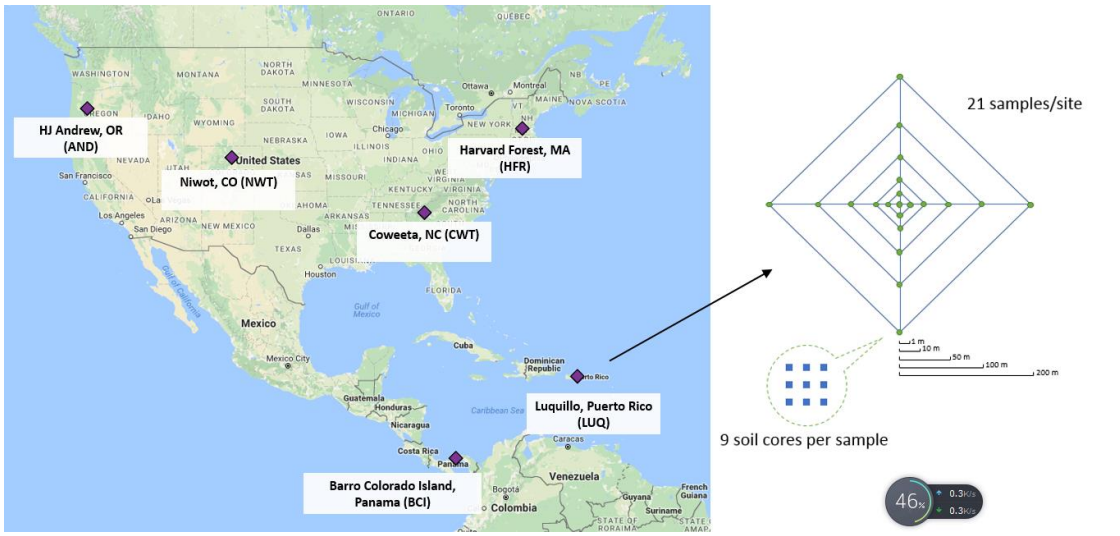


Figure 3.1 Sampling sites and sampling strategy with nested design. At each site, 21 nested samples were collected at distance of 1, 10, 50, 100 and 200 m. Nine soil cores were collected and composited in each sampling site for microbial and soil analysis. The sites information can be found at the project website: http://macroeco.lternet.edu.

3.3.2 Metadata collection

The Plant species were surveyed using a modified ‘Gentry plot’ methodology whereby five 0.1-ha Gentry plots were established by B.J. Enquist, V. Buzzard and S.*et al*. within the 25-ha plot within each site. Mean annual temperature and average annual

precipitation were calculated from hourly data collected from onsite weather stations. The soil moisture was measured by putting 1.5 g soil into 65 °C oven until a constant weight was reached. The percentage of the original weight loss after oven drying was calculated as the soil moisture content (%). Soil pH was measured in a soil suspension with a soil: water ratio of 1:2.5 (weight: volume) using a standard protocol described previously (Zhou, Deng et al. 2016). The soil C and N contents were measured by a LECO TruSpec Carbon and Nitrogen Analyzer (LECO Corporation, St. Joseph, MI) in the Soil, Water and Forage Analytical Laboratory at the Oklahoma State University (Stillwater, OK). In the same analytical laboratory, the soil NH_4^+ and NO_3^- contents were extracted from the soils with 1 m KCl and measured by a Lachat QuikChem 8500 series 2 instrument (Lachat, Loveland, CO). More detailed information about this project and metadata collection can be found at http://macroeco.lternet.edu/.

3.3.3 DNA extraction and Illumina sequencing

Soil DNA was extracted using the grinding SDS-based DNA extraction method as previously described (Zhou, Bruns et al. 1996). The quality was assessed based on spectrometry absorbance at wavelengths of 230, 260 and 280 nm (ratios of absorbance at 260/280 nm ~1.8 and 260/230 nm >1.7) detected by a NanoDrop ND-1000 Spectrophotometer (NanoDrop Technologies). DNA concentration was measured by PicoGreen using a FLUOstar OPTIMA fluorescence plate reader (BMG LABTECH, Jena, Germany).

The phasing amplicon sequencing approach (Wu, Wen et al. 2015) was used. an amplicon of 309 bp (not including the primers) in ITS2 region was targeted using the primers: gITS7F, GTGARTCATCGARTCTTTG and ITS4R,

TCCTCCGCTTATTGATATGC. To avoid extra PCR bias that could be introduced by the added components in the long primers used for PCR library preparation, a Two-step PCR was performed for ITS amplicon sequencing. Forward and reverse primers were used so that the total length of the amplified sequences remain constant. The extra bases spacers were added to the forward and reverse primer set in a complementary manner to ensure that the exact seven bases for sequencing phase shift. The primers in both direction contains the Illumina adaptor, the Illumina sequencing primer, a spacer, the ITS primer and a 12-base barcode in the reverse primer to distinguish the samples. To do the two-step PCR, the first round PCR was carried out in a 50 µl reaction containing 5µl 10× PCR buffer II, 0.5 U high-fidelity AccuPrimerTaq DNA polymerase (Life Technologies), 0.4 µM of both forward and reverse primers and 10 ng soil DNA. Reactions were performed in triplicate and the sample amplification was performed in 10 cycles, with the annealing temperature was 56°C for ITS. The triplicate products from the first round PCR were combined, purified with an Agencourt AMPure XP kit (Beckman Coulter, Beverly, MA, USA), eluted in 50 µl water and aliquoted into three new PCR tubes (15 µl each). The second round PCR used a 25 µl reaction containing 2.5 µl 10 × PCR buffer II (including dNTPs), 0.25 U high-fidelity AccuPrime Taq DNA polymerase (Life Technologies), 0.4 µM of both forward and reverse phasing primers and 15 µl aliquot of the first round purified PCR product. The amplifications were cycled 20 times following the above program. Positive PCR products were confirmed by agarose gel electrophoresis. PCR products from triplicate reactions were combined and quantified with PicoGreen.

PCR products from samples to be sequenced in the same MiSeq run (generally 3 × 96=288 samples) were pooled at equal molality. The pooled mixture was purified with a QIAquick Gel Extraction kit (Qiagen Sciences, Germantown, MD, USA) and re-quantified with PicoGreen. Sample libraries for sequencing were prepared according to the MiSeq Reagent Kit Preparation Guide (Illumina, San Diego, CA, USA) as described previously (Wu, Wen et al. 2015). First, the combined sample library was diluted to 2 nM. Then, sample denaturation was performed by mixing 10 µl of the diluted library and 10 µl of 0.2 N fresh NaOH and incubated 5 min at room temperature. A measure of 980 µl of chilled Illumina HT1 buffer was added to the denatured DNA and mixed to make a 20 pM library. Finally, the 20 pM library was further adjusted to the desired concentration (~12 pM) for sequencing using chilled HT1 buffer. The library for sequencing was mixed with a certain proportion of a Phix library of the same concentration to achieve a 10% Phix spike.

A 300-cycle v1 (for 16S ribosomal DNA, rDNA) or 500-cycle v2 (for ITS or *nifH*) MiSeq reagent cartridge (Illumina) was thawed for 1 h in a water bath, inverted 10 times to mix the thawed reagents and stored at 4 °C for a short time until use. For 16S rDNA sequencing, customized sequencing primers for forward, reverse and index reads were added to the corresponding wells on the reagent cartridge before being loaded as described previously (Wu, Wen et al. 2015). Sequencing was performed for 251, 12 and 251 cycles for forward, index and reverse reads, respectively.

3.3.4 Sequence processing and annotation

The raw ITS sequences were collected in Miseq sequencing machine in FASTQ format. First, the sequences were mapped into samples using the barcode information in each

sequence with no mismatch allowed. Then the forward and reversed reads were joined together as a single sequence using FLASH (Magoc and Salzberg 2011) program when there were at least 10bp overlap and <5% mismatches between the two reads. To further control the quality of the sequences, BTRIM (Kong 2011) was used to filter the sequences with the threshold of QC > 20 over 5bp window size. Any joined sequences with ambiguous bases or with length less than 200bp were discarded. Thereafter, U-CHIME (Edgar, Haas et al. 2011) was used to remove chimeras by searching against UNITE ITS reference dataset released on Jan 12th, 2016. Operational taxonomic units (OTUs) were clustered using UCLUST (Edgar 2010) with the identity similarity of 97%. Thereafter, the reads of OTUs were re-assigned back to their samples and a matrix with 126 samples as columns and all OTUs as rows was generated for each data set. The OTUs appeared in only one sample were considered as singletons and excluded from most of the statistical analysis.

For the classification of the ITS sequences, the representative sequences generated in the OTU clustering process were used to identify the taxonomic information for all the sequences belong to the corresponding OTUs. First, the representative sequences were searched against UNITE database (Koljalg, Nilsson et al. 2013) using BLASTn (Altschul, Gish et al. 1990) to find the closest hit with known taxonomic classification. The UNITE database released on 11.20.2016 was used as references and reference sequences without identified genus information are not included in the BLAST search. We relied on 90, 85, 80 and 75% sequence identity as criterion to assign OTUs in Genus, family, order and class level respectively. The search result with e-values > e^{-20} are not considered. We followed Index Fungorum

(www.indexfungorum.org) for genus to phylum level taxonomy of fungi as suggested in FHiTHINGS (Dannemiller, Reeves et al. 2014). The taxonomy database used in FHiTHINGS is complemented and sorted with new genera to match the up-to-date UNITE sequence. Then we used the lowest common ancestor algorithm implemented in FHiTHINGS (Dannemiller, Reeves et al. 2014) to classify these sequences. For the ambiguous classification from the lowest common ancestor algorithm, RDP (Wang, Garrity et al. 2007) ITS classifier was further used to classify these sequences with the Warcup training set provided in the RDP website. And sequences were re-classified in the phylum level with the confidence level no less than 0.5.

To assign potential function groups known for fungi community, FUNGuild (Nguyen, Song et al. 2016) was applied to find the most possible match for the OTUs. We use USEARCH (Edgar 2010) to search for most close sequences in UNITE database for each OTU based on their global similarities. The taxonomic information of the best match is assigned to each OTU and further analyzed by FUNGuild to assign functional groups based on their databases. To ensure accuracy, taxonomy information is kept when the query sequence matched to ≥93% similarity in the UNITE database.

3.3.5 Statistical methods

The OTU richness, Shannon index and Chao1 value are used to estimate the α-diversity of the fungal community from the forest soil samples. One-way ANOVA are applied to test the overall community differences in α-diversity indices (OTU richness and Shannon index) among six forest sites, and Turkey's test are used as post-hoc tests to further analyze the difference between taxonomic groups in different levels. To investigate the best environmental predictors of fungal richness (OTU number), we used

multiple linear regression in mixed models as implemented in R package “lme4”, which adds random effects to the model to account for site variations that cannot be explained by the predictors. After removing variables with collinearity, it is not surprised that the full model using all the rest environmental variables does not have random site effects (the variance of random intercepts is very close or equals to zero) upon the fungal richness, which means that the variables included in the full model explains all the variations between sites that can affect the OTU numbers. Then we used both backward selection method (stats::step function in R) and manually removing the least significant predictors, to reduce the model to best possible model consist of the best predictors. The models are chosen based on their AIC scores. The relative importance of these richness predictors is determined by the forward selection process based on adjust R squares as implemented in the “adespatial” package in R. This selection process can also help to validate the selected model by constrain the accumulated alpha value at a significant level.

The fungal community dissimilarities between different sampling sites can be used to estimate β-diversity. Bray-Curtis distance between each fungal community is used to calculate the dissimilarities. Mantel test using Pearson correlation showed that the dissimilarities using OTU abundance and OTU richness profiles are highly correlated with $r=0.9022$ and $p\text{-value}<0.001$, so we use the OTU richness dissimilarities to estimate the β-diversity between samples, to reduce the potential bias introduced by the abundance.

To investigate the distance decay relationship, we used the distance-decay curve to estimate the turnover rate of the fungal species in these forests. The turnover rate can

be estimated through the coefficient (slope) of a linear least squares regression between the log transformation of the distance and the log transformation of the similarity between the fungal microbial communities from these forest sites (Martiny, Eisen et al. 2011). This approach uses comparisons of the communities rather than the estimation of species richness in an area. To get the distance-decay curve, distances between each plot within one site is calculated directly from the nested sampling design strategy as shown in **Error! Reference source not found.**. The distance between sampling sites are transformed from the latitudinal and longitudinal coordinates using 'haversine' method implemented in package "geosphere" in R, which account for the spherical nature of earth, but ignoring the ellipsoidal effects. The total distance-decay relationship of all the 126 samples were observed as well as the distance-decay curves for the six forest sites separately. The z-score represents the species-area relationship can be also estimated from the slope of the distance-decay curve as demonstrated in the previous study (Green, Holmes et al. 2004). The distance-decay slope should be negative two times the z-score based on the definition. The samples, not the distance matrix cells were permuted 999 times to get a randomized slope distribution, and the observed slope was compared to the distribution to test for significance (Martiny, Eisen et al. 2011).

To discover the relationships between environmental factors and the fungal diversity, multiple regression models were used. The environmental factors were standardized first, and simple linear regression was performed to detect the collinearity. The plant richness/diversity and elevation data were removed from the candidate factors due to the collinearity with other factors, which means they can be retrieved from the linear combination. We used the fixed effect model ("lme4" package in R) to remove

the effect of autocorrelation within sites. It is interesting that we found the random effects of the site variable was zero or indistinguishable from zero when using different subset of predictors, which indicates the site effect can explain no more variance than the other predictor. Therefore, we only use the regular multiple linear regressions for the fungal richness prediction purpose. The environmental factors were selected using the forward and backward selection criteria based on AICs (Akaike information criterion) using the 'step' function in the "stats" package in R. After the final factors were selected to build the best prediction model, the relative importance of these components was determined using the forward selection method based on accumulated adjusted R square, as implemented in the "adespatial" package in R.

3.4 Results

3.4.1 Sequencing results

A total of 4,944,616 ITS sequences were obtained for 126 samples from six forest sites after merging the raw pair-end reads. The OTU picking analysis revealed that the sequences were clustered into 30222 OTUs after filtering low-quality and potential chimeric reads. Among these OTUs, 24.4% (7378) OTUs were singletons, which contain only one sequence across all the samples. These singleton OTUs were discarded as they are commonly considered erroneous sequences. All the samples were randomly resampled at 19727 sequences per sample and the 21954 OTUs still remaining are used for the further statistical analysis.

Across all the soil samples, most ITS sequences belonged to a small number of OTUs and the majority of the OTUs were much less abundant. For example, the top 200 (0.9%) of the most abundant OTUs covered 40.85% of all the sequences, and the top

2000 (9.1%) of the most abundant OTUs covered 83.93% of all the sequences. The most single abundant OTU (classified as a member of Ascomycota) accounted for 1.57% of all the sequences. Among the six forest sites, the majority (71.63%) of the OTUs is unique to the site, which means that they were only found in one site. Only 1912 (8.71%) OTUs were found in at least 3 sites, but these OTUs represented 64.13% of all the sequences from all the sampling sites.

3.4.2 Fungal community composition across the six forest sites

The taxonomic annotation analysis shows that the fungal communities sampled from the six forest soils in North America covers most major phyla of fungi. However, there were still 2155 (9.82%) OTUs (accounted for 1.76% sequences) cannot be classified at the phylum level neither by comparison to the annotated sequences in UNITE database at 75% similarity level, nor by the Naïve Bayesian RDP classifier with 50% confidence level. Among the fungi phyla, Basidiomycota (48.03%) and Ascomycota (47.85%) encompassed the largest proportion of the classified sequences (Figure 3.2), and the rest sequences belonged to Chytridiomycota (0.13%), Zygomycota (1.43%) and Glomeromycota (0.71%).

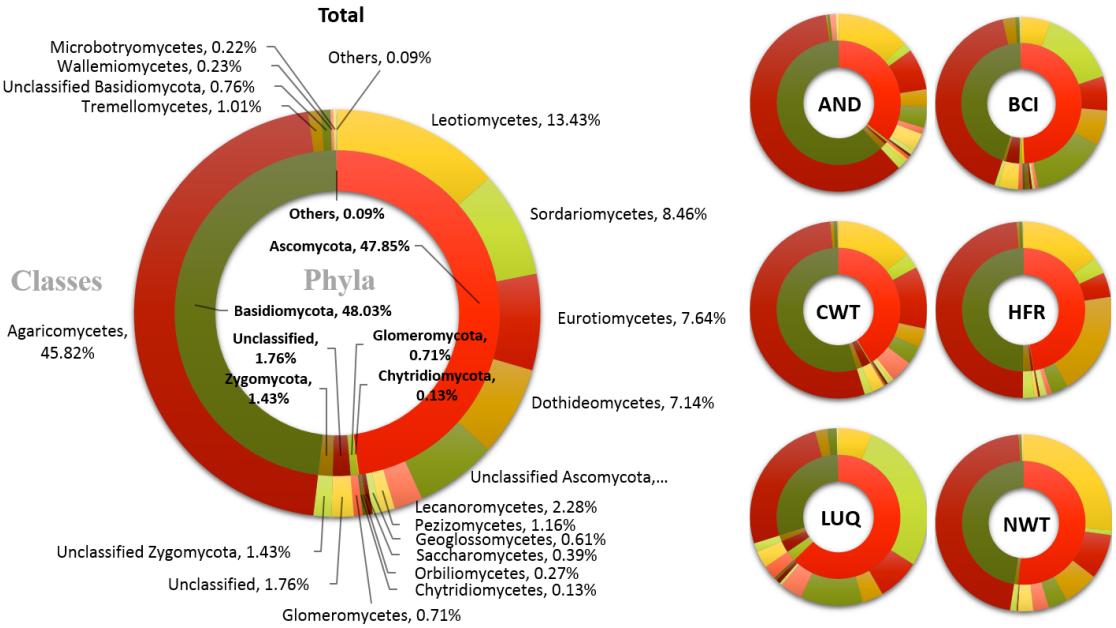


Figure 3.2 Relative sequence abundance assigned to major fungal phyla and classes. The left panel is the relative portion of all the ITS amplicon sequences collected from the six forest soils; the right panel shows the detailed taxa group distribution in each of the forest site.

The main phylogenetic fungal groups were present in all the six forest soils, but their relative portions varied across these sites. For example, the ratio of Basidiomycota to Ascomycota species abundance was highest (1.75) in AND, the temperate conifer forest, but lowest (0.47) in LUQ, one of the tropical forests. When only considering the richness, the ratio of Basidiomycota to Ascomycota OTUs was still highest (0.71) in AND, and lowest (0.42) in LUQ. Glomeromycota were also relatively more diverse in LUQ (2.72%) while Zygomycota OTU richness peaked (3.79%) in HFR, the temperate deciduous forest. Chytridiomycota accounted for a small proportion of OTU richness across six sites (0.43% ± 0.13%).

Besides the taxonomic groups, the fungal functional groups, also called ‘guild’ in FUNGuild (Nguyen, Song et al. 2016) program, also show differences in the studying sites. There are three trophic modes based on the nutrient source: symbiotroph,

pathotroph and saprotroph, which can be further categorized into more detailed ‘guilds’ as shown in **Figure 3.3(b)**. The FUNGuild pipeline assign these functional annotations based on the taxonomic classification, especially the species information, so the sequences that cannot be classified or with ambiguous species information cannot be annotated. With the current FUNGuild database, we can see that only 8% to 30% sequences as shown in **Figure 3.3(a), Figure 3.3(Figure 3.3 (c)** can be annotated with the functional groups. It is interesting that the percentage of the sequences can be annotated is decreasing as the mean annual temperature increases, which indicates that the sites with higher temperature contains more soil fungal species that have not yet been recorded or studied. This is consistent with the alpha diversity pattern, that tropical sites tend to have higher diversity than the temperate and boreal sites, as discussed in the section before. In the temperate and boreal forest sites, the most abundant guilds are the ectomycorrhizal fungi and undefined saprotrophs in the soil, which matches well with previous studies (Hogberg, Baath et al. 2003, Buee, Reich et al. 2009, Nguyen, Song et al. 2016). The relative abundances of undefined saprotroph are also high in the two tropical forest sites BCI and LUQ, while the ectomycorrhizal species relative abundances are much lower than their abundances from the other four sites.

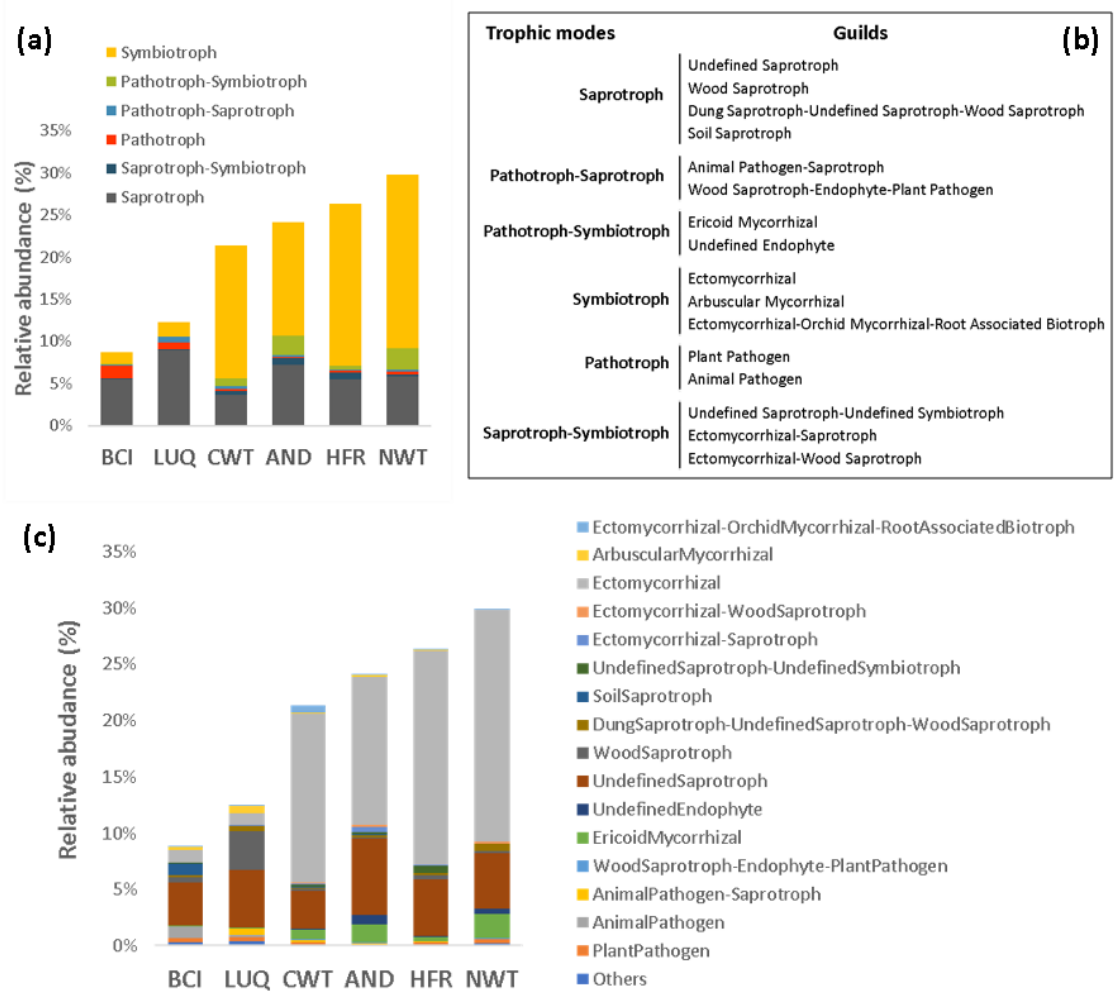


Figure 3.3 Fungal functional group distribution across the six forest sites, as defined in FUNGuild (Nguyen, Song et al. 2016). (a) Relative abundance of trophic modes in different sites; (b) relationship between the top abundant guilds and corresponding trophic modes; (c) relative abundance of guild among the sites.

3.4.3 α-diversity pattern and its drivers

The α-diversity of the fungal communities were estimated using Chao1, OTU richness and Shannon diversity index as shown in **Figure 3.4**. The ANOVA test of these three indexes confirmed that the fungal communities from the six forest soils were significantly different from each other, with all the *p*-values less than 0.001. To further investigate which communities were different, post hoc tests were used to separate the fungal communities into groups with different α-diversity indexes (**Figure 3.4**). The

rainforest BCI site had much higher α-diversity than other sites. For the Chao1 richness index, the six forest sites were separated into three groups, in which BCI from forest with the lowest latitude belonged to the group with the highest Chao1 value, while the three sites with the highest latitudes (AND, HFR, NWT) belonged to the group that had the lowest Chao1 estimations. The OTU richness followed the same pattern as the Chao1 index and were separated into the same three groups. The differences in the Shannon diversity index from the fungal communities did not show a latitude related pattern as clear as the ones showed from the richness estimators.

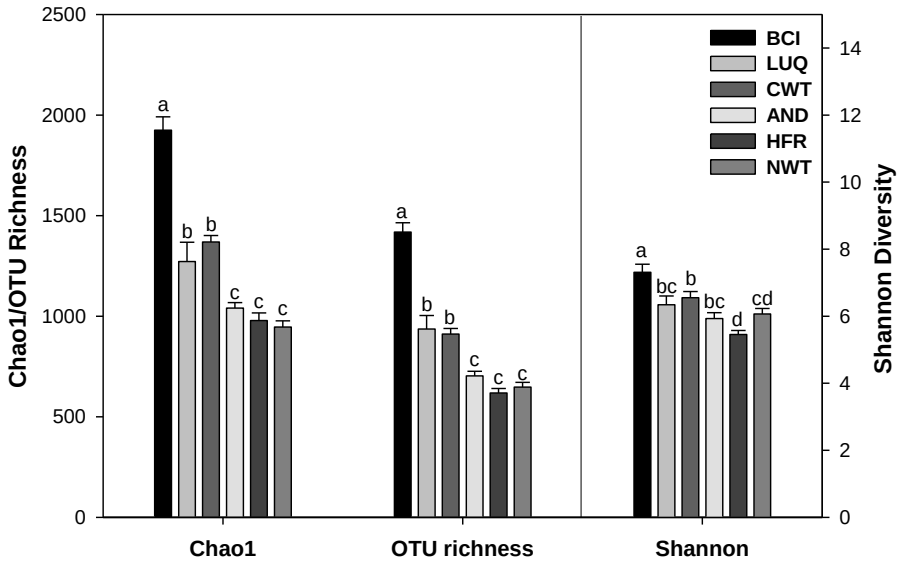


Figure 3.4. α-diversity indexes of fungal communities across six forest soils. Three indexes: Chao1, OTU richness and Shannon diversity were used for estimating the α-diversity of the soil fungal communities. The ANOVA post hoc test separate the sites into groups that have significantly different α-diversities to each other.

The distribution of each taxonomic group (phylum) across the six sites, and ANOVA test was used to test the difference among the six forest sites as shown in **Figure 3.5**. The species/OTU richness was calculated as the number of OTUs belonged to each phylum, while the relative abundance also takes the OTU abundance (sequence

number) into consideration. From the result we can see that the species/OTU richness generally decreases as the average annual temperature of the site drops. There are a few exceptions when the richness didn't strictly follow the temperature gradient, such as Ascomycota in HWT, Zygomycota in HFR, etc., but most of the phyla richness does show this clear pattern in both abundant phyla and less abundant ones. On the contrary, there are no clear pattern between the site annual temperature and the relative abundance of the phyla, and the distribution of different phyla does not show any similar patterns. The ANOVA post-hoc analysis indicates there exist significant differences between the sample sites in both the species/OTU abundance and relative abundance. But the sites belong to the same type of forests does not necessarily have similar phylum distribution in terms of OTU richness and abundance.

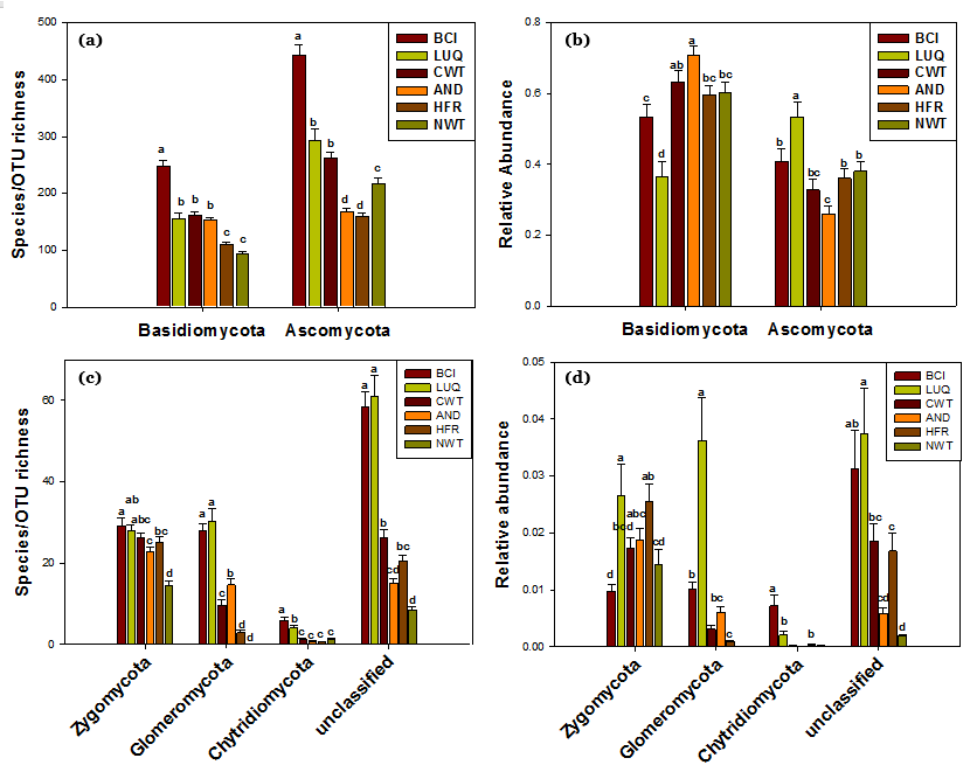


Figure 3.5 ANOVA test of the fungal phylum distribution across six forest sites. The top two figures show the OTU richness (a) and relative abundance (b) of the two most abundant phyla Basidiomycota and Ascomycota and the ANOVA post-hoc analysis results; the lower two

figures show the distribution of the rest phyla with their OTU richness (c) and relative abundance (d) and the ANOVA test result. The post-hoc analysis used Tukey HSD to test whether the means are significantly different from each other.

The best model using multiple linear regression include six variables: latitude, temperature, precipitation, soil pH, total carbon and total nitrogen in soil. In fact, the backward selection from the full set of predictors leads to a model including ammonium concentration besides the six variables in the final model. However, even the ammonium concentration does explain more variation in fungal richness as proved by R square value in the forward selection process, the selection will stop at five variables (without total carbon and total nitrogen) in the model to maintain a significant alpha value which is less than 0.05. If we use the other six variables as predictors, forward selection process will include all of them and the random site affect is still zero after removing ammonium concentration as a predictor. In fact, removing any one of the final six predictors, will results in non-zero random site effects. Therefore, we exclude the ammonium concentration in the final model and keep the other six variables. The model is significant with adjusted R square of 0.7476, with all the variables has a significant contribution as listed in **Error! Reference source not found.**. The model suggested that latitude is the top contributor, followed by precipitation, pH, total carbon, total nitrogen and temperature.

The variation partitioning analysis results (Figure 3.6) shows the explaining power of the environmental variables in the best multiple regression model. With the model, 25.24% variation of the species richness cannot be explained, while the rest of the variation can be explained by the six predictors in three groups. Soil variables include soil pH, total carbon, total nitrogen; climate variable include mean annual temperature

and precipitation; and latitude is treated as the third group to represent the locations of the fungal microbial communities. After controlling for soil and climate factors, latitude itself can attribute 2.73% to the species richness variation. When use the same set of environmental factors to predict the Shannon index (α-diversity with species abundance), 72.20% of the variation cannot be explained and climate variables (mean annual temperature and precipitation) contributes only 0.52% to the variation in the Shannon index when other factors are controlled for.

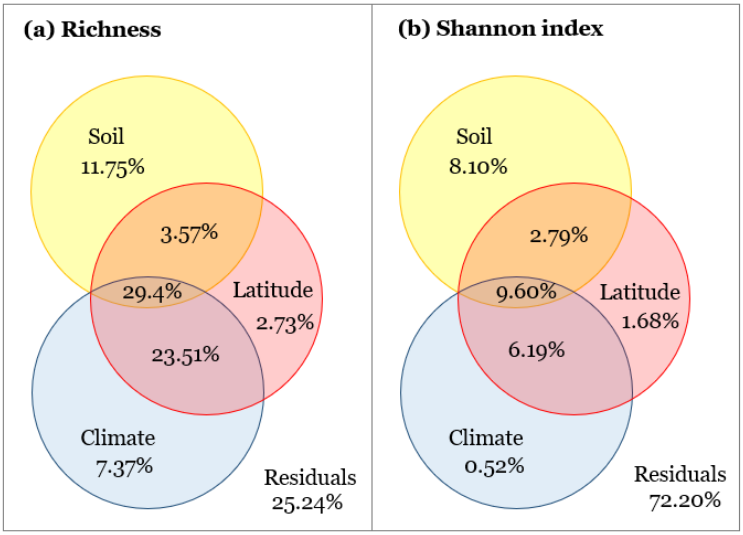


Figure 3.6 Variation partitioning analysis of fungal community (a) richness and (b) Shannon index. All the explaining variables are from the best multiple regression model. Soil variables include soil pH, total carbon, total nitrogen; climate variable include mean annual temperature and precipitation. The numbers indicate the percentage of the variation that can be explained by certain group of the factors.

3.4.4 β-diversity and distance-decay pattern

The detrended correspondence analysis reveals that the fungal communities from the same forest site were more similar and therefor tended to cluster together, while the fungal communities from the different sites were well separated in the DCA biplot (**Error! Reference source not found.**). The samples were distributed along with the

DCA1 axis as the latitude of the samples decreased, while the samples from the two tropical forest sites were well separated along the DCA2 axis. To further demonstrate the dissimilarities among the fungal communities in these six sites, three non-parametric multivariant dissimilarity analysis were performed. These dissimilarity tests all confirmed that the fungal community structures from different forest soil were significant different from each other with p-values ≤ 0.001.

The relative importance of environmental factors versus geographic distance to the fungal community similarity differed across different spatial distances (**Table 2.2** Results of the multiple regression on matrices analysis by spatial scale). Geographic distances had a strong effect at all the spatial scales we measured, from within sites to all the sites across continental scales. It is expected that the geographic distances have a larger effect at the continental scale (coefficient b=0.73), and the effect is the minimum when measured at a local scale (b=0.255, within sites). The relative importance of other environmental variables also varied by scale. Soil moisture seems has no effect on the fungal community structures at any scale, while the total soil carbon can only explain a small portion of the variation between sites. The concentration of nitrate and soil pH are important at all the scales to explain the dissimilarities among the fungal communities. Since the annual mean temperature and elevation are the same for samples from each site, there are no within-sites variances, but they do explain some of the variances exists, especially at the site level.

Table 2.2 Results of the multiple regression on matrices analysis by spatial scale

	within sites	**between sites**	**All scales**
	$R^2 = 0.255$	$R^2 = 0.428$	$R^2 = 0.651$
Ln(geographic distance)	0.338**	0.303**	0.73**
Total carbon			
Ammonium	0.134**		0.048**

Nitrate	0.103*	0.049*	0.053**
pH	0.242**	0.096**	0.071**
Moisture			
Temperature	na	0.478**	0.092**
Elevation	na		

if a partial regression coefficient is reported, the p <= 0.05, * p <= 0.01, **p<=0.001

Within the six distinct forest sites, fungal community OTU similarity decayed significantly with distance in different rates as shown in Figure 3.7 and **Error! Reference source not found.**. All the six regression coefficients are statistically different than zero, even though the absolute values are small. Permutation test results indicate that these distance decay patterns we observed cannot be achieved by random permuted samples. We also calculated taxa-area z-score from the slopes of these distance decay curves. To determine whether the site variables are the main factors affect the z-score, we computed the correlations between site variables and the z-scores. Interestingly, the site elevation has the most significant Pearson correlation (ρ) of -0.949 (p = 0.0039), while the latitude does not have a significant correlation with their z-scores (ρ = -0.407, p =0.423). The temperature is the other factor also highly correlated with the z-score (ρ = 0.755, p =0.083). When remove the effect of temperature, the partial correlation between the elevation and the z-scores is still significant (ρ = -0.878, p =0.005).

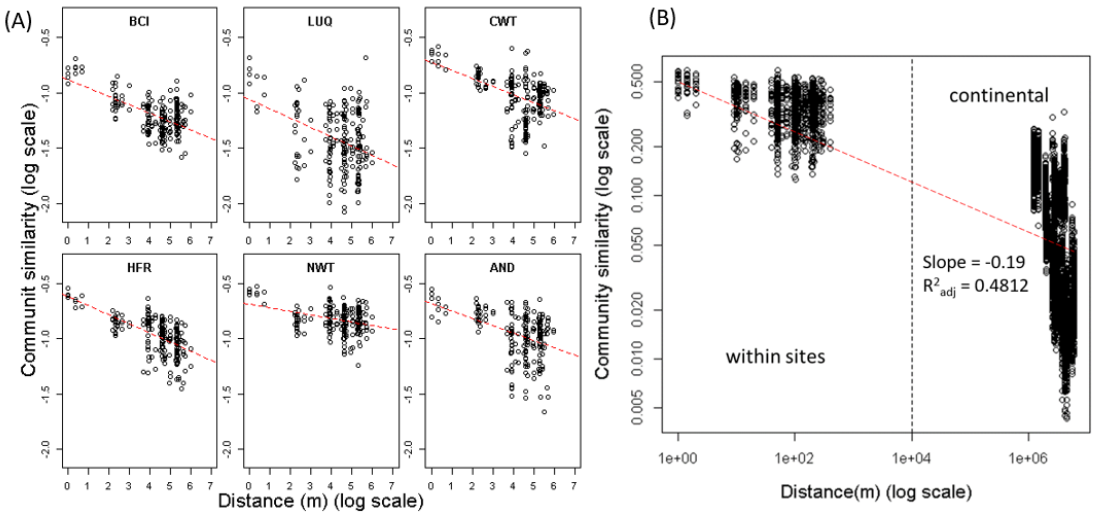


Figure 3.7 The distance-decay of similarity for microbial fungal OTUs in (A) six sites (B) all sites. The statistics of each plot and regression are listed in table 3.3.

Table 3.3 Summary statistics for the fungal OTU distance-decay in the six forest sites in North America (* denotes the slope in the linear regression model is significantly different than zero with p<0.001)

Sites	Within sites regression statistics			Permutation test
	slope	r^2	z score	p-value
BCI	-0.0764	0.382*	0.0382	<0.01
LUQ	-0.0850	0.1467*	0.0425	<0.01
CWT	-0.0730	0.2152*	0.0365	<0.01
AND	-0.0673	0.17*	0.0337	<0.01
HFR	-0.0820	0.37086*	0.0410	<0.01
NWT	-0.0317	0.1194*	0.0159	<0.01

3.5 Discussion

The fungal microbial communities have distinct community composition across the six forest sites in North America. The overlap between these six sites are extremely low in the number of OTUs, even between the same type of forests (e.g., tropical forests for BCI and LUQ). This means that the local communities how a high degree of spatial autocorrelation, while communities with longer distances have a higher level of community dissimilarity. This finding provides the support of the existence of

endemism of soil fungal species in North America, which are consistent with the results from other studies of fungal communities (Robeson, King et al. 2011, Talbot, Bruns et al. 2014). As a consequence of geographic endemism, soil fungal communities displayed a significant distance-decay pattern from local to continental scale. The z-score (from 0.0159 to 0.0425) in the distance-decay curve in this study is smaller than the z-score reported previously (Feinstein and Blackwood 2012) at site level, and it increased when the special scale become larger ($z = 0.095$). This pattern has also been observed for microorganisms and it indicates that dispersal limitation is possible for microorganisms at larger special scales, which will cause the z-score to increase. Various environmental variables such as mean annual temperature, soil pH and soil chemistry, can explain some of the differences between the distinct fungal communities, spatial distance are the major factor contributes to the diversity differences at different forest sites.

Soil fungal richness has shown to follow the latitudinal pattern, which means that the closer to the equator the fungal community is, the higher species number it has, consistent with bacteria and macro organisms. The most intuitive way to explain this pattern is that the environmental conditions do changes along with the latitude gradient, such as temperature and precipitation. It has been shown that, given the metabolic theory of ecology, temperature can better predict the taxonomic and phylogenetic distances than other environmental variables, such as soil pH (Zhou, Deng et al. 2016). This explains why the latitude, which is highly correlated with the soil temperature, are the most important factor that can be used to explain the differences in the α-diversity of the fungal communities. The fact that latitude has the most predictable power in

multiple linear regression model and it still contribute to the richness variance alone after controlling for these two variables, reveals that latitude alone is an important indicator of the fungal richness. It may be the evidence to prove the hypothesis that all the species are originated from the equator. The reason could also be that there are environmental variables that we didn't measure correlate with latitude but not with other variables that we did measure.

Generally, it is expected to find that fungal abundance along altitudinal (elevation) gradients should decrease, since the increase of elevation usually indicates a harsher environment (Margesin, Jud et al. 2009). In our study, however, elevation is not directly linked the fungal community abundance or diversity. It could due to the reason that the sites are not distributed along an altitude gradient in the local scale, therefore the environmental gradients created by the elevation differences can not be captured in such scale. But interestingly, the z-score in the distance decay curve, has shown a significant correlation with the elevation, which indicating dispersal limitation may play an important role when shaping the fungal (and/or fungal related plant) community structure along the altitude gradient.

The predictive power is only limited to the fungal species richness but not to the α-diversity with the abundance considered. As shown by the results of ANOVA test (Figure 3.4, Figure 3.5), there's no clear pattern shown between the Shannon index and the temperature gradient or latitude gradient. And the most significant explainers that can predict the fungal richness can only explain less than 30% variance of the fungal diversity measured by Shannon index (Figure 3.6). It is argued that the accurate measurement of abundance is difficult due to underlying PCR bias (Bellemain, Carlsen

et al. 2010) and the poor correlation between the amount of targeted genes and cell biomass (von Wintzingerode, Gobel et al. 1997), thus it is hard to use as an index in microbial ecology. It is also possible that the abundances of microorganism species are more likely controlled by the complicated micro-environment and are subject to more complex ecosystem dynamics involved more variables than could be measured.

3.6 Conclusions

In this study, we explore the soil fungal communities in six forest sites across North America. We have observed that the soil fungal community are diverse at the continental scale, with distinct taxonomic and functional composition. The α-diversity display a strong latitudinal gradient, which means the sites that are closer to the equator have a higher number of species. As demonstrated in the mixed linear regression model, latitude along with the mean annual temperature, precipitation, soil pH, soil total carbon, and soil total nitrogen. These seven variables can be used to predict the α-diversity of the soil fungal communities, and more than 70% variance can be explained by these variables only. Even though the plant richness is the most correlated variables with fungal species richness, which is expected due to the strong association between fungal species and plants, the plant richness is not included in the prediction model, since it can be explained by the linear combination of the other variables mentioned above. So, these environmental factors can also be used as predictors of plant species richness and it is possible that they will also affect the fungal communities indirectly through the plants, along with the direct influence upon the fungal species themselves. As for the β-diversity, the dissimilarities among the fungal communities increases significantly as the distance between the sampling sites become larger. This pattern can

be shown in the distance-decay curve, which provides a quotative way to estimate the turnover rate for the fungal species in the forest soil systems. The key drivers of the difference in fungal community composition highly depends on the spatial scale, and the geographic distance is the major contributor to explain these differences. In summary, this study of the fungal communities in the North American forest soils have shown several patterns along with the possible drivers behind them, which presents insights to the nature of soil fungal communities. These patterns are consistent with those observed in microorganisms, which seems universal to all the living organisms.

Chapter 4: Microbial Functional diversity and Ecosystem functioning

4.1 Abstract

Elucidating the relationships between biodiversity and ecosystem functioning is one of the grand challenges in ecology, particularly in microbial ecology. Microorganisms, as the most abundant and diverse group of life on earth, are involved in essential ecosystem functioning and services around the planet. Although high-throughput metagenomic technologies provide massive, rich data on studying microbial biodiversity, its importance in ecosystem processes is highly controversial. One of the main reasons for such heavy debate is the difficulty in defining microbial functional traits and their diversity. Here we developed a novel framework to characterize microbial functional diversity based on high throughput metagenomics technologies, mainly GeoChip-based functional gene arrays. We also used GeoChip to analyze groundwater microbiomes from highly contaminated wells before and after one-time Emulsified vegetable oil (EVO) injection at the Oak Ridge Field Research Center (Oak Ridge, TN). The new developed framework was used to assess microbial functional diversity changes in the groundwater microbiomes after the EVO injection. Our results indicate that comparing to gene richness and other functional indices, the functional diversity of the key gene (FTHFS) directly related to the EVO degradation is more closely linked to the actual biodegradation activities. Other genes involved in the following reduction of contaminants also showed significant correlations between their functional diversity and corresponding environmental variables. In addition, the differences in the environmental variables during this dynamic succession can explain a significant part of the differences in microbial community functional structures

constructed using functional diversity. These results suggest that the new developed functional diversity index can provide extended insights of the functional community structures and showed a closer linkage to the ecosystem functioning. Application of this framework will be helpful to understand the community assembly process and the mechanisms behind the biodiversity and ecosystem functioning (BDEF) relationships.

4.2 **Introduction**

Ecosystems are extremely dynamic systems consists of variance living organisms and the environment around them. These components interact and influence with each other and form complex interaction networks and the stability of these networks reflects the equilibrium of the ecosystems. The diversity of organisms plays crucial part to keep the structure of the networks, and the loss of such diversity can have detrimental effects upon the network stability and left the ecosystem fragile and vulnerable to changes in the environment. Thus, over the last decades, biodiversity and its response to environmental changes are central issues in ecology and for society. Microorganisms are the main engines of the Earth biogeochemistry cycles, and the changes of their biodiversity will lead to changes in the ecosystem stability and its functioning. It is generally believed that more diverse system could perform better than less diversity system in terms of ecosystem functioning due to the functional differences among various functional groups and the niche complementarity of different species. However, controversial results have been obtained (Flynn et al 2011; Cardinale et al. 2012; Nielsen et al. 2011; Zhou et al. 2015). Particularly the mechanisms underlying biodiversity-ecosystem functioning (BDEF) relationship are hotly debated (Tilman, Houston, Duffy 2008). One of the main reasons for such controversy and debates is originated from the use of different facets of biodiversity (Zhou et al. 2015).

Functional diversity is a developing concept and can be measured using various indices (Pavoine and Bonsall 2011). Among the existing mathematical frame works, Rao's quadratic approach has several advantages. First, it allows different dimensions of biodiversity (e.g., taxonomic, phylogenetic and functional diversity) within the same

mathematical framework. The Rao's method incorporates both the relative abundance of taxa and a measure of the pairwise differences between taxa (Ricotta 2005). Hence, it provides information on both functional evenness and divergence, which are two components should be included in functional diversity studies according to Pavoine and Bonsall 2011. Second, the Rao quadratic entropy approach provides a general framework for partitioning biodiversity into three components: α, β, and γ diversity (Ricotta 2005, Pavoine and Bonsall 2011). In addition, Rao's approach provides direct measure of functional redundancy (de Bello, Leps et al. 2007, de Bello, Lavergne et al. 2010), which is one of the few methods to measure functional redundancy within and among biological communities. Finally, various comparative studies suggested that this approach quite accurate (Clark, Flynn et al. 2012, Gagic, Bartomeus et al. 2015). All of these unique characteristics of the Rao's quadratic entropy index are very attractive for biodiversity analysis because it could open new perspectives to understand mechanisms shaping community assembly and the turnover along spatial, temporal and environmental gradients. Thus, in this study, we will use Rao's quadratic entropy to quantify functional diversity of a functional gene in a microbial community.

4.3 **Mathematical framework of functional diversity**

4.3.1 Functional traits and GeoChip database

For simple functional trait, individual gene in the genome can represent the presence or absence of the trait. To quantify theses functional traits of microbial community, closed format functional gene microarray can be used to measure the potential ability of corresponding functional genes. GeoChip (Tu, Yu et al. 2014) is a functional gene array contains probes targeting functional genes involves in various ecosystem functions and

ecological processes, such as carbon degradation, nitrogen cycling, stress responses, virulence, hydrogen production, etc,. It can be used as a powerful tool to monitor the functional composition and structure of microbial communities in response to different environmental conditions. In the past decade, the GeoChip has been kept up to date by updating and re-designing to accommodate continuously expanding public sequence databases. In the most recent GeoChip version (GeoChip 5.0), there are about 167,000 probes targeting more than 1,590 functional genes, which can be classified into several generalized functional categories, such as carbon, nitrogen, sulfur, phosphorus cycling, energy metabolism, organic remediation, stress response, bacteriophages, and virulence (Zhou, He et al. 2015).

4.3.2 Rao's quadratic entropy

Assume that *m* microbial communities are analyzed with high throughput metagenomic technologies such as sequencing (both shotgun and amplicon sequencing) and functional gene arrays. A total of *n* numbers of homologous functional genes (e.g., nirK, nifH, amoA, nosZ) important to ecosystem functioning are detected. Under each functional gene of interest, numerous gene sequences or probes were detected. Based on certain sequence thresholds, these individual sequences from each functional gene can be grouped together as individual operational unites (OTUs). The individual OTUs obtained by sequencing or the probes detected by hybridization could represent individual microbial genera, species or populations, depending on the taxonomic resolutions. For convenience of description below, we refer to individual OTUs or probes as individual taxa. The number of sequences of OTU or the intensity of a probe represents the taxon abundance. The sequence or hybridization data for each taxon

across various microbial communities can be tabulated as Table S1. Here, we treat individual functional genes detected as individual functional traits of a microbial community because they are important signatures for community functioning. In the following, we will describe approach on how to measure microbial functional diversity in a microbial community based on Rao's quadratic entropy.

4.3.2.1 Functional diversity

Let s_k be the number of taxa of the k^{th} functional trait (gene) detected across all communities, and x_{ikl} represent the abundance of the i^{th} taxa of the k^{th} functional trait in the l^{th} community ($i \in (1,2,\dots,s_k); k \in (1,2,\dots n); l \in (1,2,\dots m)$). Therefore, for the k^{th} gene, we have the abundance matrix $X^{s_k \times m} = [x_{ikl}]$ across all m communities. The relative abundance p_{ikl} is the proportion of the abundance of the i^{th} taxon of the k^{th} gene in the l^{th} community to the abundance of all the taxa detected for this gene in this community, which can be denoted as in equation (1), where $i \in (1,2,\dots,s_k); k \in (1,2,\dots n); l \in (1,2,\dots m)$ and $\sum_{i=1}^{s_k} p_{ikl} = 1$.

$$p_{ikl} = \frac{x_{ikl}}{\sum_{i=1}^{s_k} x_{ikl}} \quad (1)$$

Based on the Rao's quadratic entropy (Rao 1982), the functional diversity of the k^{th} gene in the l^{th} community can be calculated as

$$FD_{kl}^{\alpha} = \sum_{i=1}^{s_k} \sum_{j=1}^{s_k} d_{ijk}\, p_{ikl} p_{jkl} = 2 \sum_{i>j}^{s_k} d_{ijk}\, p_{ikl} p_{jkl} \quad (2)$$

where α denotes this is the α-diversity of funtional trait k; d_{ijk} is the pairwise dissimilarity or divergence between taxon i and j for the k^{th} functional trait. FD_{kl}^{α}

measures the functional α-diversity of the k^{th} gene in the l^{th}community, which is the average difference between any two selected taxa of the k^{th}gene in the l^{th} community.

The variance of the unbiased FD_{kl}^{α} can be estimated (Shimatani 2001) by

$$var(FD_{kl}^{\alpha}) = \frac{4}{S_k(S_k-1)}\Big[(3-2S_k)\Big(2\sum_{i>j}^{s_k} d_{ijk}\, p_{ikl}p_{jkl}\Big)^2 + (n-2)\sum_{i,j,t}^{s_k} d_{ijk}d_{itk}p_{ikl}p_{jkl}p_{tkl} + \sum_{i>j}^{s_k} {d_{ijk}}^2 p_{ikl}p_{jkl}\Big] \quad (3)$$

Therefore, gerenal significant test based on normal distribution with this variance can be used to test the difference between different microbial communities.

4.3.2.2 Partition of functional diverity (α, β and γ-diversity)

Partitioning biodiversity into different spatial components of (α, β and γ) is important to disentangle the processes and mechanisms shaping biodiverist and tis turnover (Meynard, Devictor et al. 2011). The diversity within a community is defined as α-diveristy, while the diversity between communities is usually defined as β-diversity. The overall diversity in a region, including both α and β-diversity is defined as γ-diversity (Whittaker 1960). When α and γ-diversity are known, the β-diversity can be calculated either by muliplicative ($\beta = \gamma/\overline{\alpha}$) or additive ways ($\beta = \gamma - \overline{\alpha}$) (Lande 1996). Since Lande's publication, the additive diversity partition has rapidly become a unifying framework that provides a quantitative description of the within- and between-community diversity at different levels of organization. Based on the additive definition, the partition of Rao's entropy in to α and β component has been proved both mathematically feasible and biologically meaningful (Ricotta 2005, Villeger and Mouillot 2008, de Bello, Lavergne et al. 2010).

To calculate functional γ-diversity in a region, all local communities examined are pooled as a single smapling unit. Let S_k be the total number of taxa in the region, and P_{ik} be the regional relative abundance of the i^{th} taxon for the k^{th} trait (gene), that is

$$P_{ik} = \frac{\sum_l^m x_{ikl}}{\sum_l^m \sum_{i=1}^{S_k} x_{ikl}} \quad (4)$$

Note that m is the total number of the local communities, and the difference between Eq.1 and Eq.4 is that Eq.4 combines all the local commnities as a single community. Therefore, the regional functional γ-diversity for the k^{th} trait (gene) can be defined as

$$FD_k^{\gamma} = \sum_{i=1}^{S_k} \sum_{j=1}^{S_k} d_{ijk}\, p_{ik} p_{jk} \quad (5)$$

Therefore, the additive functional β-diversity (FD_k^{β}) for the k^{th} trait is the difference between the functional γ-diversity and the average funcitonal α-diversity across all communities:

$$FD_k^{\beta} = FD_k^{\gamma} - \overline{FD_k^{\alpha}} \quad (6)$$

As argued in (Villeger and Mouillot 2008), to avoid negative functional β-diversity, in the equation (Eq. 6), the average of the should be defined as the weighted average of the α-diversity, where the weight (w_{kl}) should be the proportion of the k^{th} trait's abundance associated with the l^{th} community in the whole region (m is the community number in the region):

$$\overline{FD_k^{\alpha}} = \sum_l^m w_{kl}\, FD_{kl}^{\alpha} \quad (6)$$

the weight w_{kl} is calculated as:

$$w_{kl} = \sum_i^{s_k} x_{ikl} / (\sum_l^m \sum_i^{s_k} x_{ikl}) \text{ and } \sum_l^m w_{kl} = 1 \quad (6)$$

4.3.2.3 Corrected functional α, β and γ diversity

When the distances between taa is all equal to 1, meaning that each taxa is uniq from each other, the Rao's entropy becomes Gini-Simpson index. Due to the biased estimation of species diversity indeices, β-diversity estimated using Gini-Simpson's formulation is always underestimated (de Bello, Leps et al. 2007, Jost 2007), Also, neither the additive or multiplicative estimation of β-diversity is ecologically meaningful when applied to Gini-Simpson's index (Ricotta and Szeidl 2009). Such bias can be resolved by introducing the Rao's equivalent number of species as a 'corrected' form for the diversity index, where c means 'corrected':

$$FD_{kl}^{\alpha,c} = 1/(1 - FD_{kl}^{\alpha}) \quad (7)$$

Similarly, the corrected form of functional γ-diversity is:

$$FD_{k}^{\gamma,c} = 1/(1 - FD_{k}^{\gamma}) \quad (8)$$

And the corrected form of functional β-diversity is:

$$FD_{k}^{\beta,c} = FD_{k}^{\gamma,c} - \overline{FD_{kl}^{\alpha,c}} = FD_{k}^{\gamma,c} - \sum_l^m w_{kl} FD_{kl}^{\alpha,c} \quad (9)$$

4.3.2.4 Functional redundancy

Generally, taxonomic diversity (TD) in a microbial community is estimated based on phylogenetic markers such as 16S rRNA or 18S rRNA and ITS (Zhou, He et al. 2015). If gene markers are capable of reflecting the differences among individual populations or taxa, they can be used to measure taxonomic diversity of different functional

assemblages or groups. Previous studies indicated that many functional genes important to biogeochemical cycling can provide species/strain level resolution (Tiquia, Wu et al. 2004, Zhou, He et al. 2015), and hence functional genes can be used as markers for measuring taxonomic diversity of various functional guilds. Determining the linkage between taxonomic linkage between taxonomic and functional diversity is critical to understanding the relationship between biodiversity and ecosystem functioning, but they are poorly understood in ecology (Micheli and Halpern 2005). It is generally believed that changes in functional diversity rather than taxonomic composition affect the resistance and resilience of ecological community structure (Bellwood, Hoey et al. 2003). Functional redundancy, i.e., the number of taxonomically distinct taxa which perform similar ecological functions, is critical concept in ecology. However, it is difficult to define functional redundancy in microbial ecology due to the lack of connections between taxonomy/phylogeny and functions. Rao's quadratic approach provides a direct estimation of functional diversity and the functional redundancy (FR) can be defined as the difference between taxa diversity and functional diversity (Pillar, Blanco et al. 2013).

When $d_{ijk} = 1$ for a all $i \neq j$, the FD_{kl}^{α} in Eq (2) becomes Gini-Simpson diversity (D), which can used as the estimation of taxonomic diversity that was measured by the functional gene maker:

$$TD_{kl}^{\alpha} = \sum_{i=1}^{S_k} \sum_{j \neq i}^{S_k} p_{ikl} p_{jkl} \tag{9}$$

$$= \sum_{i=1}^{S_k} p_{ikl} \times \sum_{j=1}^{S_k} p_{jkl} - \sum_{i=1}^{S_k} {p_{ikl}}^2 = 1 - \sum_{i=1}^{S_k} {p_{ikl}}^2$$

Therefore, from Eq (2) and Eq (9), the functional redundancy of the functional redundancy can be defined as (de Bello, Leps et al. 2007):

$$FR_{kl} = \ TD_{kl}^{\alpha} - FD_{kl}^{\alpha} \tag{10}$$

Or in the more recenlty defined form as (Ricotta, de Bello et al. 2016):

$$FR_{kl} = \ (TD_{kl}^{\alpha} - FD_{kl}^{\alpha})/TD_{kl}^{\alpha} \tag{11}$$

4.3.2.5 Community level functional diversity and redundancy

The definitions in the above sections are all focus on a single functional trait (gene), the k^{th} trait in the l^{th} community. Generally, a community l has different types of functional traits, such as traits related to nitrification, denitrification, nitrogen fixation, carbon decomposition, and sulfate reduction. The overall functional α-diversity of the l^{th} community can be expressed as

$$FD_{l}^{\alpha} = \ \sum_{k=1}^{n} q_{kl} \, FD_{kl}^{\alpha} \, , \tag{12}$$

where q_{kl} can be 1 (for unweighted) or the proportion of abundance of the k^{th} functional trait in the l^{th} community (for weighted):

$$q_{kl} = \ \sum_{i}^{s_k} x_{ikl} / (\sum_{k}^{n} \sum_{i}^{s_k} x_{ikl}) \text{ , and } \sum_{k}^{n} q_{kl} = 1 \tag{13}$$

4.3.3 Quantifying distances between taxa

In the definition (Eq 2) of functional diversity FD_{kl}^{α}, the relative abundance p_{ikl} can be interpreted from the signal intensity of functional gene arrays, while d_{ijk}, the pairwise dissimilarity between taxon i and j for the k^{th} functional trait, needs to be provided additionally. Unlike animal and plants, there are not much information available for

functional traits from individual taxa due to the unculturable nature of most microorganisms. Fortunately, the DNA sequencing technology can provide rich phylogenetic information about the degree of the relatedness of taxa in a microbial community.

For correct spatial partition (α, β, γ) of biological diversity, one of the primary mathematical requirements is that the estimated quadratic diversity should be concave (Ricotta 2005), that is the total diversity in a set of communities should be greater than or equal to the weighted or average diversity within the communities (Lande 1996). Then the total taxa diversity in a pooled set of communities can be partitioned into additive components of within-community and between-community diversity. The Rao's quadratic entropy is proved to be concave if the taxa distance is Euclidean (Ricotta 2005). The Euclidean distance matrix can be simply obtained by taking the elementwise square root of the distance matrix extracted from a phylogenetic tree, rooted or unrooted (de Vienne, Aguileta et al. 2011). In addition, the Rao's method should be estimated based on ultrametric distance to assure the index reaches its maximal value when all the taxa are retained (Pavoine, Ollier et al. 2005). A distance matrix $D = [d_{ij}]$ is ultrametric if and only if $d_{ii} = 0$, $d_{ij} \geq 0$ and $d_{ij} \leq \max(d_{it}, d_{tj})$, for all taxa i, j, t. The distance matrix obtained from a phylogenetic tree with all tips are equidistant from the root, such as trees generated using UPGMA clustering algorithm (Sokal 1958), is ultrametric (Pavoine, Ollier et al. 2005). Using ultrametric distances in Rao's entropy diversity will avoid the situation that the index reaches its maximum when only several extreme taxa exist while others are absent, which is usually the case when using just Euclidean distances (Botta-Dukat 2005, Pavoine, Ollier et al. 2005).

Finally, the pairwise distance (d_{ijk}) should vary from 0 to 1 (Botta-Dukat 2005, de Bello, Lavergne et al. 2010). If $d_{ijk} = 1$ for all taxa, then Rao's quadratic approach becomes Gini-Simpson diversity (Botta-Dukat 2005). Thus having d_{ijk} ranges from 0 to 1 have the advantage for generalized framework for different diversity indexes.

Two major types of phylogenetic approaches can be used to estimate taxon divergences: distance-based methods and tree-based methods. Among distance-based methods, one could simply use pairwise sequence dissimilarity (1 – similarity) to quantify the differences between two taxa, which is also called p-distance (Nei 2000). However, direct estimation of sequence similarity based on nucleotide sequences generally underestimates the differences among different organisms due to mutation saturation (i.e., some of the nucleotide positions may have experienced multiple substitution events) (Van de Peer 2009). Thus, the pairwise dissimilarity among taxa can generally be corrected based on different evolutionary models, such as Jukes-Cantor distance (Jukes TH 1969), Tajima-Nei distance (Tajima and Nei 1984), Tamura 3-parameter distance (Tamura 1992), and Tamura-Nei distance (Tamura and Nei 1993). These estimated phylogenetic distances are not Euclidean distances, but they can be transformed into Euclidean distances by simply taking element-wise square root of the distance matrix (Legendre and Anderson 1999, de Vienne, Aguileta et al. 2011).

However, distance-based approaches could not catch enough phylogenetic information because phylogenetic trees are not used so that the relationships among multiple species (> 3) are not clear. A phylogenetic tree, a branching diagram or "tree" showing the inferred evolutionary relationships among various biological species or other entities, is the best way to catch the relationships among different taxa. Thus, we

will also use phylogenetic tree-based approaches to estimate the divergences among different taxa. Three major approaches are often used to construct phylogenetic trees, including distance-matrix methods (UPGMA, neighbor joining), maximum parsimony, and maximum likelihood. While the tree constructed by UPGMA is ultrametric, others are not. In an ultrametric phylogenetic tree, i.e. a tree in which all tips have the same distance to the root, and the distances extracted from an ultrametric tree is ultrametric distances.

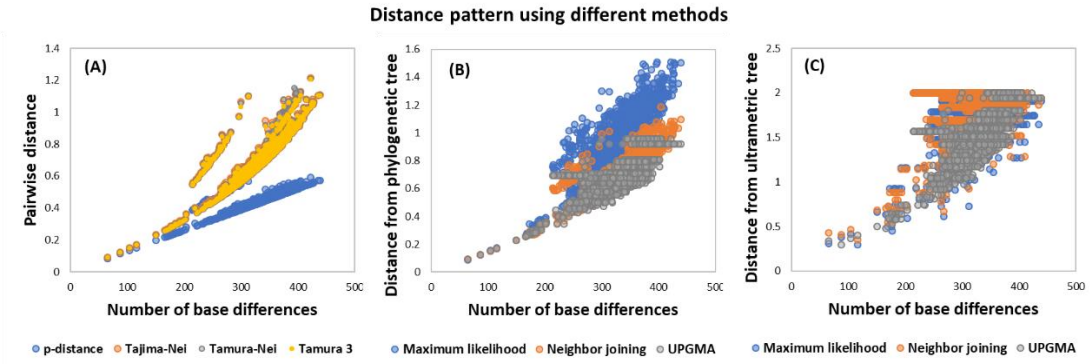


Figure 4.1 Distance pattern using different distance methods. (A) Direct distance with evolutionary models (B) Phylogenetic distances extract from trees constructed using three method: maximum likelihood, neighbor joining, UPGMA. (C) ultrametric distances from time-corrected phylogenetic trees.

There are two ways to calculate pairwise taxa distances based on phylogenetic trees. (i) Node numbers-based methods: The d_{ijk} can be defined as the number of internodes from the species level to the lowest level of the phylogenetic tree in which a common ancestor of tax i and j share (Guiasu and Guiasu 2010, p 710-711). The estimated phylogenetic distance should be transformed as Euclidean distance and be standardized to vary from 0 to 1. (ii) Branch lengths-based methods. Similarly, several ways can be used to estimate pairwise phylogenetic distances between species based on branch lengths. The first is to directly calculate branch lengths based on the output files generated by various phylogenetic programs (Gaith et al 1992, Allen et al. 2007; Chao,

2010). The phylogenetic distances based on branch lengths can then be transformed into Euclidean distance and standardized to vary from 0 to 1. Another is to render the non-ultrametric trees to ultrametric tree by relaxing global clock assumption or by post hoc tree transformation via penalized likelihood rate smoothing (Sanderson 2002). Then pairwise cophenetic distances can be estimated based on ultrametric tree using the method from the ape package (Paradis et al. 2004) (Fig. 1). The cophenetic distances can then be standardized to vary from 0 to 1 to represent the phylogenetic divergence among taxa (de Oliveura et al. 2014). However, based on the original algorithm, cophenetic distance approach may loss distance information as the finer level. Theoretically, the former (directly estimating branch length) is preferred. In addition, one could also use divergence time as estimating the species difference as described previously (Hardy and Senterre 2007). The divergence time can be estimated based on phylogenetic tree (Sanderson 2002).

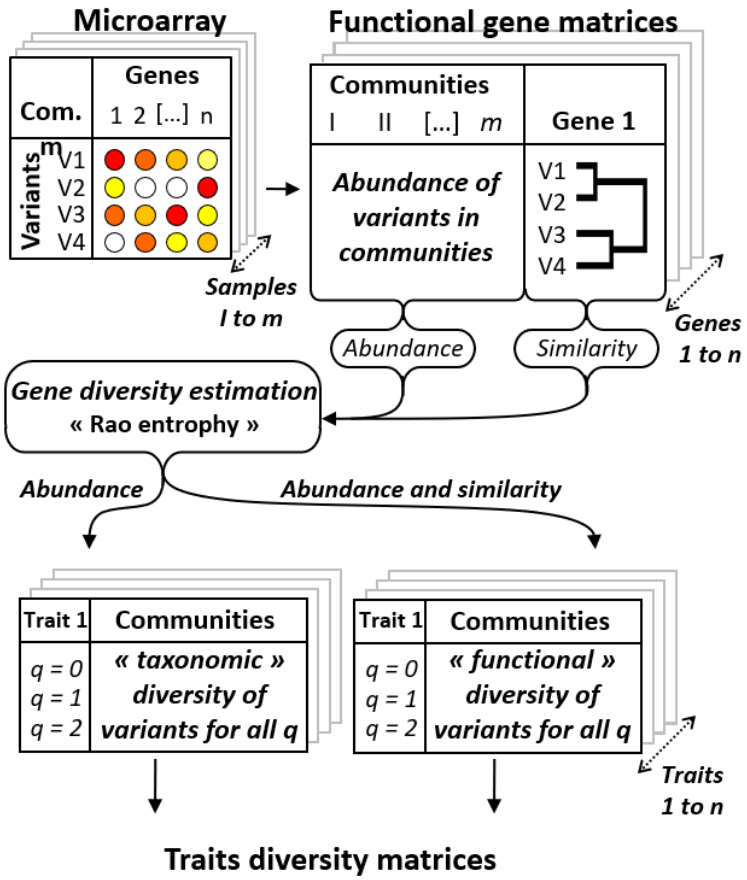


Figure 4.2 Conceptual framework of functional diversity profiles from GeoChip data

4.3.4 Pipeline construction

First, DNA sequences their corresponding protein sequences extracted from GeoChip database from three versions: GeoChip3, GeoChip4 and GeoChip5 (Table 4.1). For each version, separated framework are constructed, since the sequences have been changing dramatically during the these GeoChip development processes, and earlier versions probably contain outdated or later updated sequences, but they are still meaningful to analyze studies using these GeoChip versions. There are different sub-versions for each GeoChip versions, and the most manufactured sub-versions are selected to cover the core genes relating to the essential ecological processes by design.

For the purpose of diversity analysis, genes have less than 10 sequences are not considered in the framework. After the extraction, the DNA sequences are corrected for open reading frames by comparison to their protein sequences using FrameBot program. The DNA sequences cannot be corrected (by inserting or deleting bases) will be discarded from the framework. To obtain the distances among taxa (represented by DNA sequences here), the DNA sequences for each gene are aligned using MUSCLE in MEGA7 as protein encoded sequences. The sequences are manually checked to remove short sequences (cannot overlap with others) and noisy sequences (from homologs, or annotation error), which will ensure successful alignments. For many functional genes, sequence diversities are very high, and to make sure there are enough common overlaps between sequences to allow alignment, pair-wise deletion with at least 95% site coverage are used during the alignment.

Table 4.1 The functional genes and categories included in the framework

Gene Category	GeoChip3		GeoChip4		GeoChip5	
	Gene No.	Probe No.	Gene No.	Probe No.	Gene No.	Probe No.
Carbon Cycling	32	3576	83	25103	99	19164
Nitrogen	15	2981	18	7405	28	5846
Organic Remediation	86	6703	98	8879	74	10133
Phosphorus	3	566	2	892	6	3099
Sulfur	4	1083	15	4603	24	4108
Other	1	1123	2	65	58	9892
Antibiotic resistance	10	1118	10	1534		
Energy process	2	94	4	508		
Metal Resistance	30	3892	40	9557		
Bacteria phage			21	445		
Bioleaching			15	275		
Fungi function			64	3737		
Soil benefit			20	1559		
Soil borne pathogen			23	497		
Stress			40	9414	86	25155
virulence			10	1433	89	10943

Electron transfer					8	659
Metal Homeostasis					98	40472
Secondary metabolism					41	3604
Virus					54	2290
Total	**183**	**21136**	**465**	**75906**	**665**	**135365**

Three phylogenetic tree construction algorithms (neighbor joining, maximum likelihood, UPGMA) are used to build functional gene trees. In these algorithms, the building process started from one or a couple of sequences and then adding other sequences as new tips gradually. Even strategies can be used to select the initial sequences and decide the adding orders of the rest sequences, the high diversity of the functional sequences always leads to ties among sequences to be chosen at a certain point. Therefore, the initial orders of the input sequences are crucial for the final gene tree structures. In other words, change of the orders of the sequences will lead to different tree structure in all the three methods. Therefore, to find the most reasonable trees, we shuffle the alignment 100 times to obtain 100 trees for each gene using every method and extract the distances between taxa from these trees (using *cophenetic* function in R). We assume that each tree structure is reasonable to some extent, so the most reliable distances should be the one that are most correlated with the average of 100 distance matrix extracted from the 100 trees with different structures. Using this strategy, one final distance matrix can be generated for each gene using each tree construction method. The distance matrix is normalized to 0 and 1 in order to be suitable for Rao's entropy calculation. The final functional diversity is calculated using Rao's entropy (*divc* function in R package 'ade4') for each gene from the GeoChip microarray data and the distance matrix between taxa from this gene. The GeoChip

profile provides the abundance for each taxon in different communities, and the distance matrix provides the dissimilarity measures among these taxa.

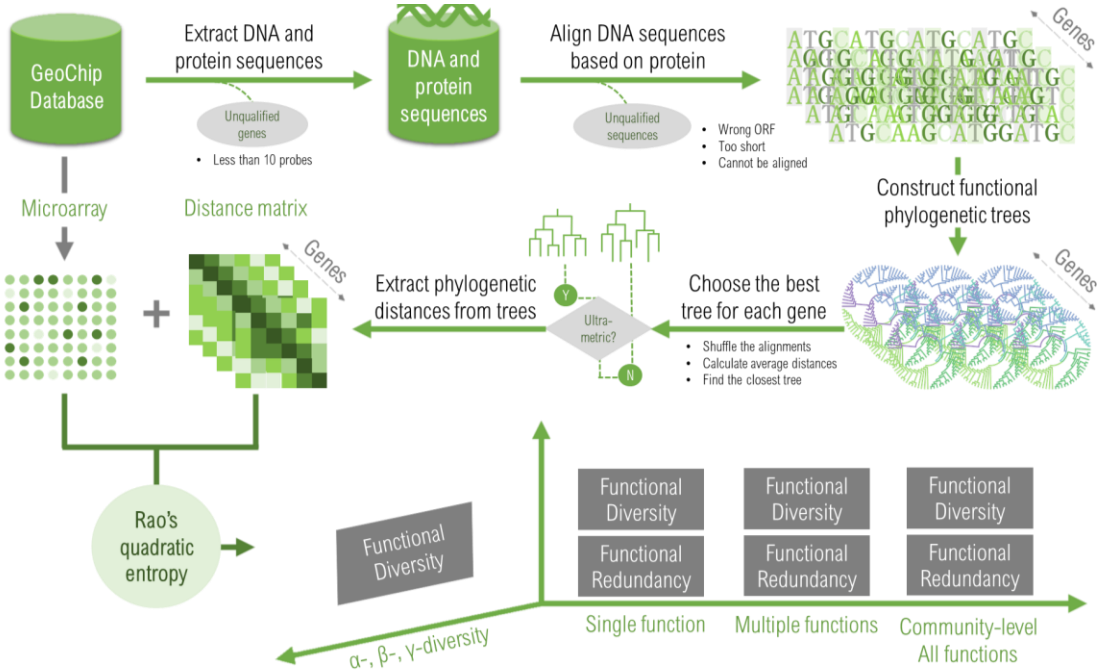


Figure 4.3 Development of functional diversity framework and databases

4.4 **Applications and results**

4.4.1 Groundwater dataset

To test the new index of functional diversity using Geochip, we use a dataset from high contaminated (U(IV), Fe(III), NO_3^-, SO_4^{2-}) groundwater samples, where EVO (emulsified vegetable oil) was injected from three injection wells, and samples were collected before the injection and after 4, 17, 31, 80, 140, 269 days from one upgradient well (W8) as control well and seven downgradient wells (W1-W7) as monitor wells. More detailed site information and sampling processes are described previously (Zhang, Wu et al. 2015). EVO amendment has been shown to promote U(VI) reduction efficiently in this site, and stimulate the aquifer microbial community with the changes

in the composition (Gihring, Zhang et al. 2011) and related function response (Chourey, Nissen et al. 2013, Zhang, Wu et al. 2015). The GeoChip analysis of these groundwater samples has shown a dynamic succession of key genes/groups involved in EVO degradation, and reduction of NO_3^-, SO_4^{2-}, and some other heavy metal contaminations. The functional diversity index proposed in this study was used to investigate the functional structure change after the EVO injection stimulates the microbial community in this groundwater system. The analysis was based on GeoChip 3.0, including 181 functional genes in seven gene categories (Table 4.1), which covers ~80% of all the proves detected in the experiment. Ultrametric distance between each probe (taxa) extracted from a time-dated phylogenic tree were used as the dissimilarity measures in the Rao's entropy definition.

Key geochemical variables were changed significantly during the 9-month monitoring period after EVO ejection. Before injection (Day 0), the groundwater samples contained a considerable amount of NO_3^- (0.2150.16 mM), SO_4^{2-} (1.14±0.11 mM) and U(VI) (8.06±2.33 μM), but the concentration of acetate was below detectable (FigureS3). After EVO amendment, substantial acetate production was observed in the seven downgradient wells, along with the obvious reduction of NO_3^-, SO_4^{2-}, U(VI), Fe(III) and Mn(IV) were also detected comparing to the concentrations in the control well. Among these, the concentrations of acetate in control well were remained undetectable during the whole observation period, which indicates that the acetate observed in the monitor wells after ejection of EVO was from the presumed biodegradation of EVO. Then the acetate will stimulate the growth of microbes that participate in the reduction process of U(VI), NO_3^-, SO_4^{2-}, Fe(III) and other metal

contaminations. Therefore, the observed acetate concentration will not accumulate and can be used as an indicator of EVO degradation process carried out by the microbial community in the site.

4.4.2 Linking functional diversity to ecosystem functions

One of the key genes that are involved in the degradation of EVO is FTHFS (also known as *fhs* and encodes for formyltetrahydrofolate synthetase), which involved in acetogenesis for acetate production. Correlations between this gene and the concentration of acetate detected in the monitoring wells after EVO injection are shown in Figure 4.3. The functional richness (probe numbers) and functional abundance (summed probe signal intensity) of FTHFS gene is not significantly related to the production of acetate ($p > 0.05$), while the Shannon, Gini-Simpson and functional diversity calculated in our new framework showed significant correlations with the concentration of acetate. Among the diversity indices showed strong relationship to the biodegradation of EVO, functional diversity has the highest correlation ($R^2 = 0.625$, $p = 0.021$).

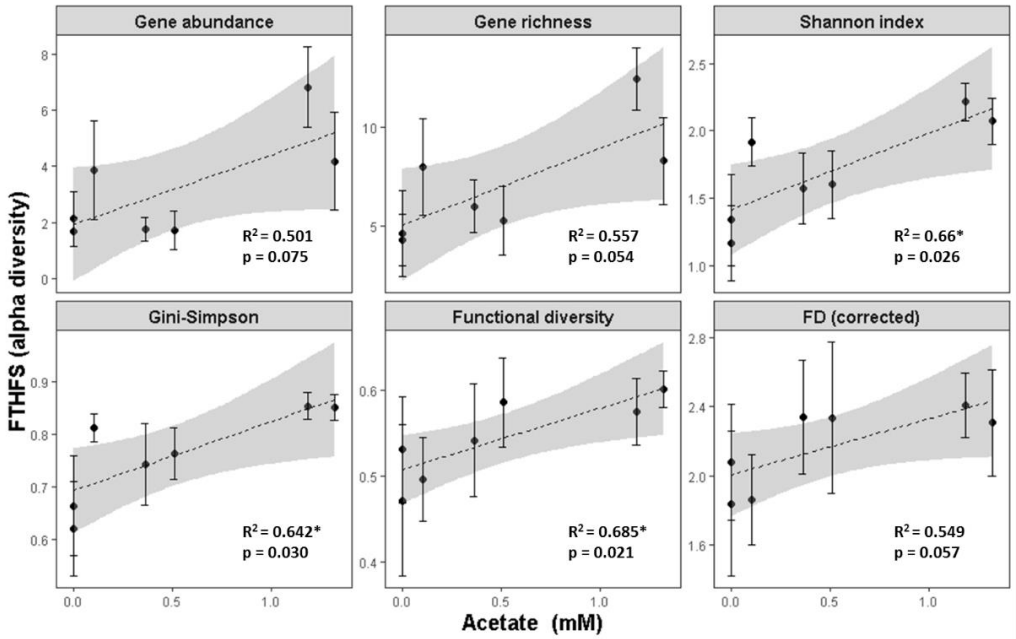


Figure 4.4 Linear relationship between function (acetate concentration) and FTHFS gene α-diversity indices (gene abundance, gene richness, Shannon index, Gini-Simpson index, functional diversity calculated in this paper and corrected functional diveristy). For each sample, gene abundance is calculated as the sum of all the probe log-transformed signal intensity; gene richness is the total number of probes detected. FD (corrected) is the corrected version of functional diversity, which is calculated as 1/(1-FD)

To further explain the correlation between functional traits and ecological process (in our case, FTHFS functional diversity and EVO biodegradation), the functional indices were standardized into same scale and plotted together along the EVO degradation progress over time (**Figure 4.5**). The gene abundance and gene richness showed almost the same trend, except when the gene abundance dropped more quickly than the gene richness after one month of the EVO injection. Both indices began to increase at the earliest time point (Day 4) that were monitored and peaked at Day 17, where they began to drop gradually and returned to almost the same level at the end (Day 269). Interestingly, the functional diversity of FTHFS gene didn't increase immediately as the gene richness and abundance did, but it dropped at the beginning and reached the lowest point at Day 4. After 4 days, the FTHFS functional diversity

kept increasing until Day 31 and then began to decrease. This change of functional diversity of FTHFS was highly consistent with the change of its presumed function: acetate production. The concentration of acetate only began to increase dramatically after Day 4 and also reached its highest level at Day 31. There was a second peak at Day 140 for both FTHFS functional diversity and acetate production, which probably due to other unreported input of organic carbon source into the system or other environmental variable changes that can accelerate the biodegradation process.

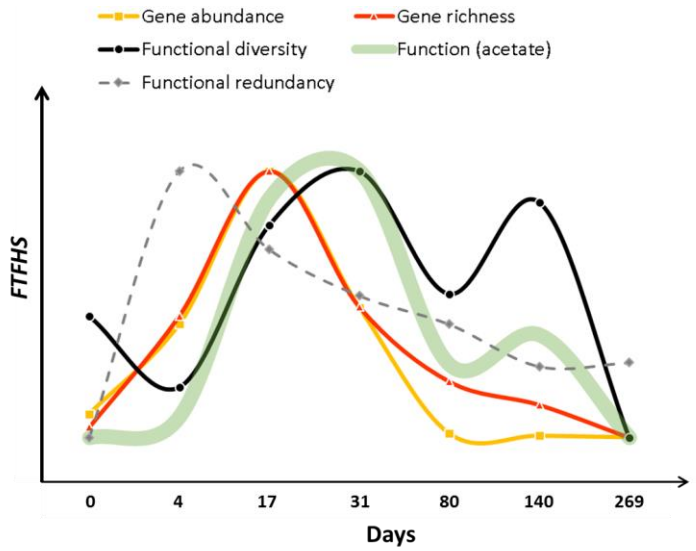


Figure 4.5 FTHFS gene diversity indices and function (acetate concentration) changes along time. All the indices were standardized to fit the same scale (0 to 1).

From this figure, the functional diversity of FTHFS gene shows a stronger linkage to the ecological process it participates in, comparing to the gene abundance and richness. At the beginning stage (from Day 0 to Day 4), the addition of new carbon source stimulates the microbial functional response by increase the corresponding gene richness capable of utilizing it, but also allows species with a narrower functional range to outgrow others, leading to a decrease in functional diversity. At this stage, the EVO degradation process starts but the efficiency is not high. After Day 4, the functional

diversity starts to increase along with the gene richness and abundance, indicating used-to-be rare species have adapted to the EVO and starts to grow better and participate in the biodegradation process. At this stage, the function carried out by FTFHS gene also increases with all the indices, since all the conditions are in favor of the degradation process. When the EVO availability declines as time goes by, the gene richness and abundance also decrease, while the functional diversity does not get affected at first (Day 17) until the resources cannot support for the minimum species richness to hold the diversity (after Day 31). The fact that acetate production peaks around the same time as the functional diversity reaches highest level, but not gene richness and abundance, shows that not all the gene richness or abundance are involved in the process. As for the functional redundancy (gene similarity), with more specified functional traits are selected at the beginning stage of disturbance, the gene similarity reaches its highest level at Day 4. Then the gene similarity decreases as time goes by, but still higher at the end of the monitor period than before EVO ejection. Detailed probe signal intensity profile is shown in Figure 4.6. The probes are listed in the order of the ultrametic phylogenetic tree constructed based on their corresponding sequences. The pattern observed here explained the functional diversity of FTHFS gene change over time. There are only four probes/taxa that are significantly correlate with the acetate production when averaging their signal intensity by the time point. The abundance of the most correlate taxa ($df = 5$, $p = 0.004$) is low, while the abundance of the dominate taxon is not correlated with the acetate production, which to some level demonstrated that functional taxa acting in a complementary way might be the reason behind the linkage between the microbial functional diversity and their functions.

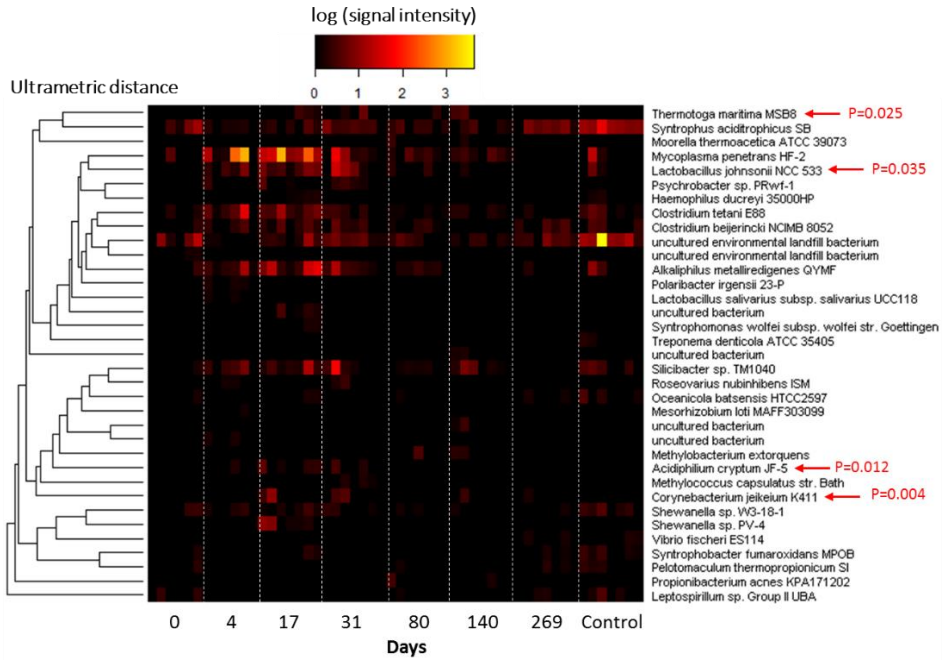


Figure 4.6 Heatmap of detailed signal intensity of FTHFS probes included in GeoChip3.0. The left panel shows the ultrametric tree, from which the distances used to calculate the functional diversity is extracted. The right panel are the species names for the probes detected in this experiment. The red arrows point at the four probes/taxa has significant correlation with acetate concentration, with p-values of the correlation listed behind.

Besides FTHFS, the correlations between key functional gene/groups and corresponding microbial processes that take place after EVO ejection are listed in Table 4.2. As mentioned above, FTHFS gene encoding for formyltetrahydrofolate synthetase involved in acetogenesis, shows strong correlations with the EVO degradation process (acetate concentration), where the functional diversity is most correlated index listed. EVO amendment also stimulate genes involved in the sequential reduction of NO_3^-, SO_4^{2-}, Fe(II) and U(VI), and other heavy metal ions which all co-exist in this groundwater system. The concentration of these electron acceptors can be also used as a measure of ecological functions (Jax 2005). For nitrogen cycling, key genes involved in nitrate reduction are evaluated for their relationship with the NO_3^- concentration. Strong correlations were found from these functional traits including genes related to reduction from nitrate to nitrite (*narG* and *napA*), nitrite reduction (*nirK/S*), assimilatory reduction

of nitrate (*nasA*, *nirA/B*). Sulfate reducing bacteria (SRR) are frequently detected in groundwater system with U(VI) contamination, which are believed to play important roles in the bioremediation in such sites. The genes (*dsrAB*) encoding dissimilarity sulfate reductase also showed strong relationships between their diversity and sulfate concentration. Energy metabolism genes, such as cytochromes are demonstrated to be involved in U(VI) reduction (Shelobolina, Coppi et al. 2007), and the functional diversity of cytochromes are significantly related to the U(VI) level, while the richness of another gene (hydrogenases) responsible for transfer H_2 to cytochromes and later to U(VI) shows strong correlations to the U(VI) level as well.

Table 4.2 Correlation between ecological functions and related genes [a]

Function	Gene	Gene Category	Gene abundance	Gene richness	Shannon (H')	Gini-Simpson (D)	Functional diversity	FD (corrected)
Acetate	FTHFS (*fhs*)	Acetogenesis	0.708	0.746	**0.813***	**0.801***	**0.828***	0.741
NO_3^-	*narG*	Denitrification	-0.458	-0.742	-0.751	-0.657	-0.424	-0.427
	nirK		-0.148	-0.618	-0.581	-0.421	**-0.864***	**-0.861***
	nirS		-0.24	**-0.789***	**-0.852***	**-0.832***	**0.887****	**0.923****
	norB		-0.017	-0.695	-0.714	-0.618	-0.564	-0.724
	nosz		-0.54	-0.748	-0.76*	-0.683	-0.416	-0.404
	napa	Dissimilatory N reduction	-0.581	-0.749	**-0.789***	**-0.848***	**-0.9****	**-0.804***
	nrfa		-0.476	-0.631	-0.54	-0.409	-0.185	-0.167
	nasa	Assimilatory N reduction	0.5	-0.636	**-0.756***	**-0.76***	**-0.8***	**-0.846***
	nirA		0.0456	-0.446	-0.344	-0.272	-0.584	-0.668
	nirB		0.0712	-0.743	**-0.796***	**-0.803***	**-0.834***	**-0.866***
SO_4^{2-}	*dsrA*	Sulfite reduction	-0.484	**-0.846***	**-0.862***	**-0.824***	-0.732	-0.722
	dsrB		-0.212	-0.735	**-0.841***	**-0.865***	**0.777***	**0.781***
U	*cytochrome*	Energy process	-0.611	-0.669	-0.662	-0.581	**-0.811***	**-0.81***
	hydrogenase		-0.644	**-0.759***	-0.628	-0.521	-0.412	-0.404

[a] Pearson correlation between ecological function and gene diversity is measured using the averaged gene indices and chemical concentrations at each observation time point. Bold font indicates significant correlation with * $p<0.05$, ** $p<0.01$

Functional diversity shows an overall tighter linkage to the chemical concentrations over other indices such as gene richness, Shannon diversity and Gini-Simpson index. When correlations are detected using more than one diversity indices, the functional diversity usually showed more significant relationships (for example, *nirS*, *napA*), and it also can capture the relationship that cannot be captured by the other

indices (*nirK*, *cytochrome*). It is interesting that some functional diversity showed opposite relationships with corresponding functions, comparing to other diversity measures, such as *nirS* and *dsrB* (Figure S4), which indicates the presumed assumption that higher functional diversity will possess function potential should be considered carefully for different functional trait under different environment conditions. There are several possible reasons. One is that the function measurement, such as SO_4^{2-} concentration, cannot accurately reflect the microbial functional activities, so the linkage might be biased. Another possible reason is that most ecosystem functions, such as sulfate reduction, are complex functions that rely on multiple functional traits (genes) to accomplish, where single gene and function correlations might not hold.

4.4.3 Shifts of the overall functional structures of microbial communities

For traditional GeoChip analysis, probe signals were used to represent microbial community functional structure for each sample. In our new framework, given the functional diversity calculated for each gene, functional profile of microbial community can be expressed in a more concise and informative way. To test whether there are substantial shifts in the functional structures of the microbial community before and after EVO injection, different functional indices and three different non-parametric multivariate statistical tests are used. The functional indices selected construct community functional profiles include probe-based indices (probe signal intensity) and gene-based indices (gene richness, abundance, Shannon index and functional diversity). Three statistical tests are: analysis of similarity (ANOSIM); non-parametric multivariate analysis of variance (Adonis); and multi-response permutation procedure (MRPP). Three methods showed practically the same pattern

of the community dissimilarities for each functional index tested, so only the PERMANOVA (Adonis) results are shown in Table S6. Based on these tests, the functional community structures differed substantially after EVO injection using both probes signal intensity and gene diversity indices (Table S6). When using probe signal intensity to present community functional structure, only samples from Day 31, Day 80 and Day 140 showed no significant dissimilarities from each other, while all the other samples are statistically different from each other. When using gene-based functional profiles, more similarities among samples are found (e.g., Day 4 vs Day140, Day 17 vs Day 31), and test results using gene-diversity-based indices (Shannon index and function diversity) are almost identical (except for Day 140 vs Ctrl).

Table 4.3 Mantel test of correlation between differences in microbial functional structures and the differences in environmental variables

Functional index	Mantel			
	n [a]	Distance methods [b]	ρ [c]	p
All probes	12987	Bray-Curtis	0.0685	0.095
FD probes [d]	10670	Bray-Curtis	0.0711	0.090
Gene abundance	187	Euclidean	-0.0223	0.613
Gene richness	187	Euclidean	0.0768	0.131
Shannon index (H')	187	Euclidean	0.0887	0.093
Gini-Simpson diversity (D)	187	Euclidean	0.1086	0.053
Functional diversity	187	Euclidean	0.1398	**0.010**
FD (corrected)	187	Euclidean	0.1075	**0.044**

[a] n is the number of probes or genes that represent functional unit in the functional profiles
[b] when calculate community distance based on GeoChip probe signals, Bray-Curtis dissimilarity is used to account for missing values; Euclidean distance is used for other gene-based indices
[c] Spearman rank correlation (ρ) is used
[d] FD probes are the probes selected into the functional diversity framework, where probes must belong to genes with more than 10 probes, and also the sequences used to design the probe are of high quality and well aligned with other sequences belong to the same gene

Mantel test is used to further determine whether the differences observed in the microbial functional structures are correlated with the change of geochemical

variables during the EVO degradation process (Table 4.3). The geochemical variables used are pH, and concentrations of acetate, Cl, Ag, Al, Ba, Ca, Cr, Ga, K, Mg, Sr, Zn, NO_3^-, Fe(II), Mn(II), U(VI), and SO_4^{2-}. To deal with large amount of missing data in GeoChip probe signals, Bray-Curtis similarity distances were used for probe-based index when calculating the functional distance between two microbial communities. For gene-based indices, the information of each gene is evaluated and summarized, so the index values are continuous and rarely contain missing data and Euclidean distance method is used to represent the dissimilarities among community structures. Among all the indices tested, only community structures represented by the functional diversity and corrected functional diversity showed significant correlations with the 18 geochemical variables ($p = 0.010$ and 0.044). When the community functional structure is divided in to different functional categories, the correlations between each division of the community functional structures and environment are listed in Figure S8. The Mantel test results indicate that, when using other indices, such as probe signal intensity or gene richness, the functional structures of microbial community are more likely shaped by other factors other than the environmental factors provided in the test, even though these variables represent the major changes occurred during the EVO degradation process. Multiple regression on distance matrices analysis (MRM) was applied to show the relative importance of each geochemical variable to the microbial community structure represented by functional diversity (Table S7). The best predictor of the community functional structure is U(VI) level ($R^2 = 0.079$, $p = 0.001$), followed by SO_4^{2-} ($R^2 =$

0.033, p = 0.001), acetate (R^2 = 0.019, p = 0.026), Fe(II) (R^2 = 0.020, p = 0.033), Al (R^2 = 0.011, p = 0.016) and Ga (R^2 = 0.017, p = 0.011).

4.5 **Discussion**

Generally, functional gene abundance and diversity can explain more of the ecosystem function, comparing the taxonomic structure of microbial community (Graham, Knelman et al. 2016). In GeoChip-based analysis, the functional structure is traditionally constructed from individual taxon/probe level (Tu, Yu et al. 2014, Xue, Yuan et al. 2016) and functional diversity is represented by the gene richness (probe number). While the finer resolution (probe level) can provide more detailed information of community structures, it also brings unexpected variances that may not contribute to the relationship with the ecosystem functioning (Table S6). The functional diversity index defined in this study aggregates related information from individual taxon and provide a higher-level depiction of the functional structure of the whole community. The results showed that the whole functional profile of microbial community represented by gene-level functional diversity, comparing to the probe-level community structure, has stronger linkage to the environmental variables, which can also be interpreted as ecosystem functions in certain context (Jax 2005). The gene-level aggregation of functional diversity also enables the partition of α and β diversity (Lande 1996, Ricotta 2005, Villeger and Mouillot 2008, de Bello, Lavergne et al. 2010), which can provide a different perspective of understanding the assembly mechanism of the functional structure of microbial communities.

When define functional diversity, if individual functional units annotated with the same gene name are considered the same in terms of function potential, the

functional diversity of this gene will become gene richness. However, differences in the gene sequences leads to probable different protein structures, and thus possible different mechanisms to carry out their functions. It is generally believed that more similar sequences lead to more similar functions. There are also significant differences when very similar protein function in different species, and sometimes even in the same organism. To better discriminate these differences can improve the prediction of their ability to function in a ecosystem together. Distances extracted from the phylogenetic trees of gene sequences can reflect such differences accompanied by their evolutionary histories. Such differences in gene can sometimes explain the various response when encounter disturbance and changes in the surrounding environment, which defines the term 'response diversity' (Mori, Furukawa et al. 2013). More phylogenetically related genes should response similarly, but lateral gene transfer can obscure this pattern, that is the reason why the functional diversity cannot completely represent functional response diversity.

It is generally believed that higher functional redundancy means the higher ecosystem stability in terms of its functions. Functioning of an ecosystem includes various processes and services, and functional redundancy can be observed in two ways. One is that when additional taxa added to the system and the function numbers (functional diversity) do not change, then there is function redundancy, in other words, the added taxa is functionally redundant. The other is that when losing a member or members, the system keeps the same functioning, then the system is functionally redundant. Both scenarios can be tested and proved when communities with different taxonomic diversity profiles have similar functional diversity profiles. An example from

human oral and fecal microbial communities (part of the NIH Human Microbiome Project data), shows that given tremendously diverse 16S profiles, the function profiles of these communities are remarkably similar (Lozupone, Stombaugh et al. 2012), indicating the existence of functional redundancy in the human microbiota. The functional diversity proposed by this work can serve this purpose to investigate the functional redundancy in terms of whole system functioning when comparing the community, the taxonomic diversity and functional diversity profiles. But careful conclusions should be made when the generality of functional redundancy has been challenged, since the species role changes in different environment, which can result in drastically different biodiversity and ecosystem function relationships (Fetzer, Johst et al. 2015).

When considering different ecosystem functions or processes individually, it is possible that a microbial community is functionally redundant in one or several functions while not redundant in the others. In our framework, we defined functional redundancy based on the genetic similarity of the genes carrying out this function, which means, given certain distribution of individual genes fulfill the function, the more similar the genes are, the higher is the functional redundancy, since losing individuals become less significant in terms of the genetic potential to achieve this function. This is the reason when the redundancy of FTHFS is high, decreasing gene richness did not lead to a direct decrease in the EVO degradation process (Figure 4.5). Studying single function redundancy is necessary when the scope of the whole ecosystem functioning is hard to define or when specific ecosystem process is the research interest. However, for complex ecosystem function that requires many genes, such as photosynthesis and

methanogenesis (Martiny, Treseder et al. 2013), it is hard to measure the functional redundancy using this definition.

4.6 **Conclusion**

This study provides a framework to detect ecologically related functional traits represented by genes for microbial community and calculate the functional diversity by combining the functional richness and phylogenetic signals contained in these traits. The application of this functional diversity framework to the groundwater microbial communities with EVO amendment shows that the functional diversity has a strong linkage to the corresponding ecosystem functions and can be a powerful to investigate the functionally assembly of the microbial community under different conditions. The functional diversity can be partitioned into α and β diversities and offer more insights of community differences and the potential mechanisms behind these differences. Along the functional diversity index, functional redundancy can also be defined and can be used to evaluate if a simple function trait is redundant in the system in terms of the genetic similarity of the corresponding genes that carry out this function. In summary, the functional diversity defined in this study can construct the functional profile of microbial communities with more information, which can may provide a stronger linkage to the ecosystem functions and help researchers to understand this linkage better.

Chapter 5: Summary and output

This dissertation has contributed to the field of microbial diversity in serval ways. **First**, we proposed a new phasing amplicon sequencing approach (PAS) was developed to conquer the issue that low-base-diversity caused during the Illumina sequencing process. This method adding diversity to the sequencing targets by shifting sequencing phases among different community samples via adding various numbers of bases (0–7) as spacers to both forward and reverse primers. Our results show that the PAS method substantially ameliorated the problem of unbalanced base composition. improved the sequence read base quality (an average of 10 % higher of bases above Q30). The PAS method also effectively increased raw sequence throughput (~15 % more raw reads) and significantly increased effective reads (9–47 %) and the effective read sequence length (16–96 more bases) after quality trim at Q30 with window 5. In addition, the PAS method reduced half of the sequencing errors (0.54–1.1 % less). Combined with two-step PCR amplification of the PAS method effectively ameliorated the amplification biases introduced by the long-barcoded PCR primers. The developed strategy is robust for 16S rRNA gene amplicon sequencing, and potentially for other gene markers important to the ecosystem functional processes. **Second**, a data analysis pipeline for amplicon sequences has been established to serve the research communities. The pipeline provides the most commonly used programs to process amplicon sequencing data for genes such as 16S rRNA, ITS, *nifH*, and other genetic markers. The pipeline was based on Galaxy platform, which provide a user-friendly interface makes code-free analysis of the amplicon sequencing data possible. The pipeline has already been set up

and kept running for several years and get involved in dozens of projects from more than 200 users. The related publications are listed in the end of this section. **Third**, a practical application of amplicon sequencing investigated the biodiversity pattern of microbial fungal communities in six North American forests soils, which adds more insights to the global fungal biogeographic distribution patterns. In this part, the soil fungal samples were collected from six forest sites across a wide range of latitudes in North America with a nested design in each site. The compositions of fungal communities are distinct from each other across six forest sites. The main drivers of alpha diversity of fungi in forest soil is latitude, along with the mean annual temperature, precipitation, soil pH, soil total carbon, and soil total nitrogen. These seven variables can be used to predict the α-diversity of the soil fungal communities, and more than 70% variance can be explained by these variables only. As for the β-diversity, the dissimilarities among the fungal communities increases significantly as the distance between the sampling sites become larger. The distance-decay curve explains this pattern and indicate that the turnover rates of the fungal species are different in the local and continental scales. We further proved that, the key drivers of the difference in fungal community composition highly depends on the spatial scale, and the geographic distance is the major contributor to explain these differences. **Finally**, we provide a new framework to quantify microbial functional diversity based on Rao's entropy using GeoChip (a high-throughput functional gene array), and the phylogenetic distances between each probe is considered in the calculation. α- and β- diversity can also be investigated from this index, which extends the understanding of functional diversity pattern into different temporal or spatial scales. We applied this functional diversity

framework to study the dynamic changes over a 9-month period of microbial communities in a contaminated groundwater system (with U(VI), SO_4^{2-}, NO_3^-, etc.,) after a one-time EVO (emulsified vegetable oil) amendment. The results show that the new defined functional diversity index is not only a better indicator of ecosystem functions when only single function is considered, but also a more appropriate index to represent the whole microbial functional structure, which shows more interactions to the ecosystem it belongs to. This framework also enables the comparison of the functional structures between different microbial communities from various studies, as long the GeoChip version is the same.

Listed below are the papers published or manuscripts which were in relation to this dissertation:

Chen, Casey, Chris Hemme, Joan Beleno, Zhou Jason Shi, Daliang Ning, **Yujia Qin**, Qichao Tu, *et al.* "Oral Microbiota of Periodontal Health and Disease and Their Changes after Nonsurgical Periodontal Therapy." *The ISME journal* 12, no. 5 (2018): 1210.

Cheng, Lei, Naifang Zhang, Mengting Yuan, Jing Xiao, **Yujia Qin,** Ye Deng, Qichao Tu, *et al.* "Warming Enhances Old Organic Carbon Decomposition through Altering Functional Microbial Communities." *The ISME journal* 11, no. 8 (2017): 1825.

Deng, Jie, Yunfu Gu, Jin Zhang, Kai Xue, **Yujia Qin**, Mengting Yuan, Huaqun Yin, *et al.* "Shifts of Tundra Bacterial and Archaeal Communities Along a Permafrost Thaw Gradient in Alaska." *Molecular ecology* 24, no. 1 (2015): 222-34.

Deng, Ye, Zhili He, Meiying Xu, **Yujia Qin**, Joy D Van Nostrand, Liyou Wu, Bruce A Roe, *et al.* "Elevated Carbon Dioxide Alters the Structure of Soil Microbial Communities." *Applied and environmental microbiology* (2012): AEM. 06924-11.

Deng, Ye, Daliang Ning, **Yujia Qin**, Kai Xue, Liyou Wu, Zhili He, Huaqun Yin, *et al.* "Spatial Scaling of Forest Soil Microbial Communities across a Temperature Gradient." *Environmental microbiology* 20, no. 10 (2018): 3504-13.

Deng, Ye, Ping Zhang, **Yujia Qin**, Qichao Tu, Yunfeng Yang, Zhili He, Christopher Warren Schadt, and Jizhong Zhou. "Network Succession Reveals the Importance of Competition in Response to Emulsified Vegetable Oil Amendment for Uranium Bioremediation." *Environmental microbiology* 18, no. 1 (2016): 205-18.

Gu, Yunfu, Joy D Van Nostrand, Liyou Wu, Zhili He, **Yujia Qin,** Fang-Jie Zhao, and Jizhong Zhou. "Bacterial Community and Arsenic Functional Genes Diversity in Arsenic Contaminated Soils from Different Geographic Locations." *PloS one* 12, no. 5 (2017): e0176696.

He, Jinzhi, Qichao Tu, Yichen Ge, **Yujia Qin**, Bomiao Cui, Xiaoyu Hu, Yuxia Wang, *et al.* "Taxonomic and Functional Analyses of the Supragingival Microbiome from Caries-Affected and Caries-Free Hosts." *Microbial ecology* 75, no. 2 (2018): 543-54.

Hemme, Christopher L, Qichao Tu, Zhou Shi, **Yujia Qin**, Weimin Gao, Ye Deng, Joy D Van Nostrand, *et al.* "Comparative Metagenomics Reveals Impact of Contaminants on Groundwater Microbiomes." *Frontiers in microbiology* 6 (2015): 1205.

Herath, Anjumala, Boris Wawrik, **Yujia Qin,** Jizhong Zhou, and Amy V Callaghan. "Transcriptional Response of Desulfatibacillum Alkenivorans Ak-01 to Growth on Alkanes: Insights from Rt-Qpcr and Microarray Analyses." *FEMS microbiology ecology* 92, no. 5 (2016): fiw062.

Liang, Yuting, Yuji Jiang, Feng Wang, Chongqing Wen, Ye Deng, Kai Xue, **Yujia Qin**, *et al.* "Long-Term Soil Transplant Simulating Climate Change with Latitude Significantly Alters Microbial Temporal Turnover." *The ISME journal* 9, no. 12 (2015): 2561.

Lin, Lu, Houhui Song, Qichao Tu, **Yujia Qin,** Aifen Zhou, Wenbin Liu, Zhili He, Jizhong Zhou, and Jian Xu. "The Thermoanaerobacter Glycobiome Reveals Mechanisms of Pentose and Hexose Co-Utilization in Bacteria." *PLoS genetics* 7, no. 10 (2011): e1002318.

Penton, C Ryan, Caiyun Yang, Liyou Wu, Qiong Wang, Jin Zhang, Feifei Liu, **Yujia Qin**, *et al.* "Nifh-Harboring Bacterial Community Composition across an Alaskan Permafrost Thaw Gradient." *Frontiers in microbiology* 7 (2016): 1894.

Wen, Chongqing, Liyou Wu, **Yujia Qin**, Joy D Van Nostrand, Daliang Ning, Bo Sun, Kai Xue, *et al.* "Evaluation of the Reproducibility of Amplicon Sequencing with Illumina Miseq Platform." *PloS one* 12, no. 4 (2017): e0176716.

Wu, Liyou, Chongqing Wen, **Yujia Qin**, Huaqun Yin, Qichao Tu, Joy D Van Nostrand, Tong Yuan, *et al.* "Phasing Amplicon Sequencing on Illumina Miseq for Robust Environmental Microbial Community Analysis." *BMC microbiology* 15, no. 1 (2015): 125.

Xue, K, MM Yuan, ZJ Shi, **Y Qin**, Y Deng, L Cheng, L Wu, *et al.* "Tundra Soil Carbon Is Vulnerable to Rapid Microbial Decomposition under Climate Warming. Nat Clim Chang 6: 595–600." 2016.

Xue, Kai, Mengting M Yuan, Zhou J Shi, **Yujia Qin**, Ye Deng, Lei Cheng, Liyou Wu, *et al.* "Tundra Soil Carbon Is Vulnerable to Rapid Microbial Decomposition under Climate Warming." *Nature Climate Change* 6, no. 6 (2016): 595.

Xue, Kai, Mengting M Yuan, Jianping Xie, Dejun Li, **Yujia Qin**, Lauren E Hale, Liyou Wu, *et al.* "Annual Removal of Aboveground Plant Biomass Alters Soil Microbial Responses to Warming." *mBio* 7, no. 5 (2016): e00976-16.

Zhang, Ping, Zhili He, Joy D Van Nostrand, **Yujia Qin**, Ye Deng, Liyou Wu, Qichao Tu, *et al.* "Dynamic Succession of Groundwater Sulfate-Reducing Communities During Prolonged Reduction of Uranium in a Contaminated Aquifer." *Environmental science & technology* 51, no. 7 (2017): 3609-20.

Zhou, Aifen, Zhili He, **Yujia Qin**, Zhenmei Lu, Ye Deng, Qichao Tu, Christopher L Hemme, *et al.* "Stresschip as a High-Throughput Tool for Assessing Microbial Community Responses to Environmental Stresses." *Environmental science & technology* 47, no. 17 (2013): 9841-49.

Zhou, Jizhong, Ye Deng, Lina Shen, Chongqing Wen, Qingyun Yan, Daliang Ning, **Yujia Qin**, *et al.* "Correspondence: Reply to 'Analytical Flaws in a Continental-Scale Forest Soil Microbial Diversity Study'." *Nature communications* 8 (2017): 15583.

Appendix A Supplementary Figures

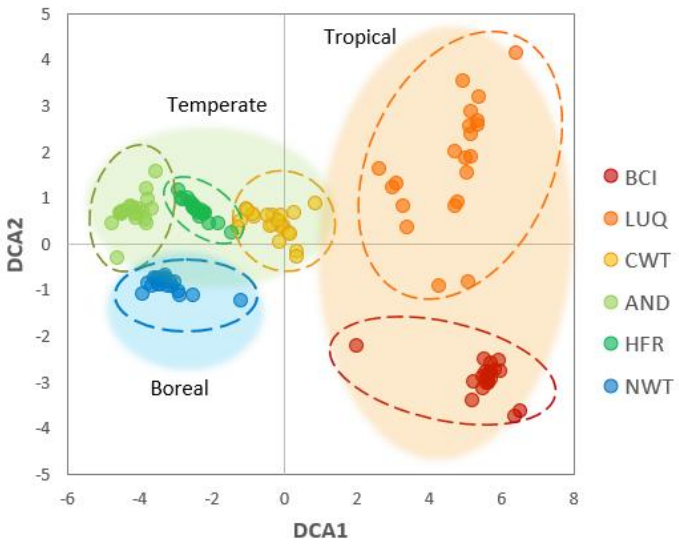


Figure S1. Detrended correspondence analyses (DCA) for fungal microbial communities in the six forest sites, including two tropical forest, three temperate forest and one boreal forest sites. These communities are clearly separated and tend to cluster by the types of the forest.

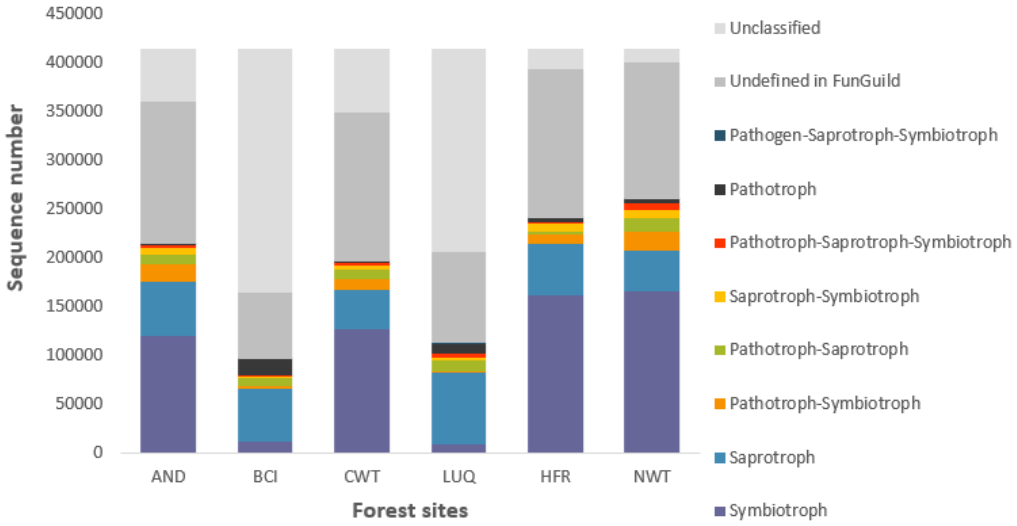


Figure S2. The sequence distribution of fungal microbial community across six forest sites based on different trophic modes.

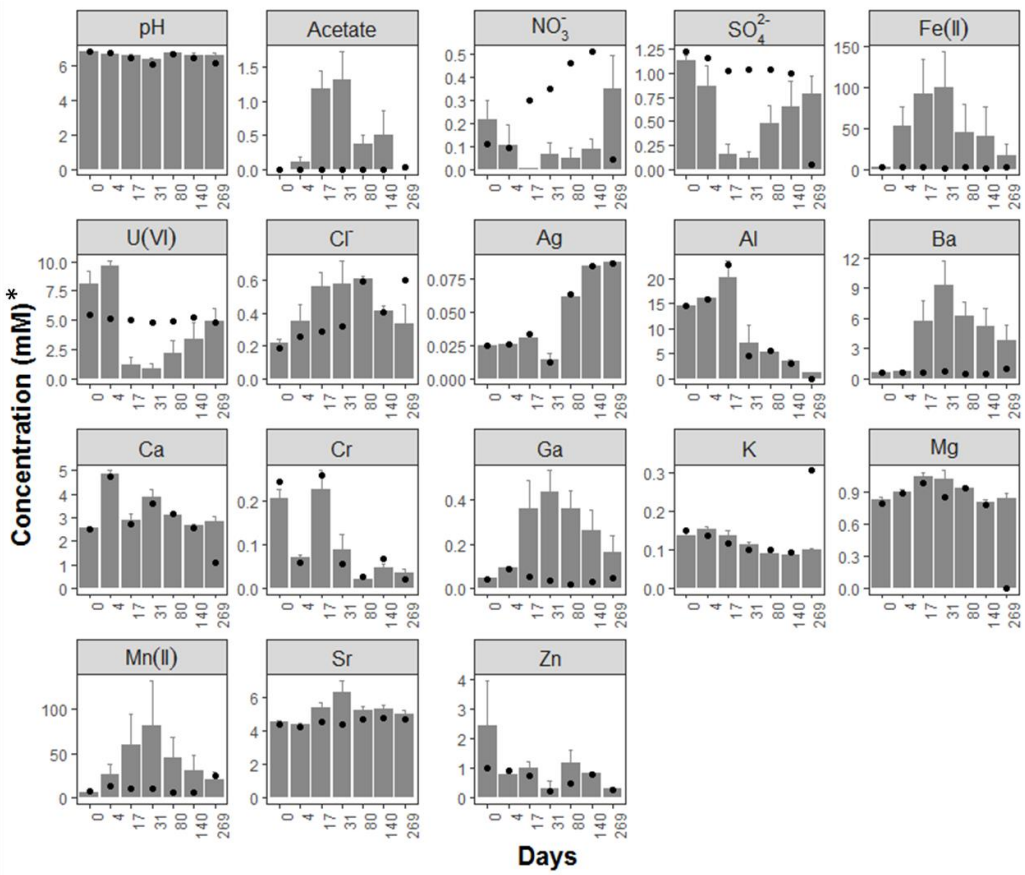


Figure S3. Changes of geochemical variables during the 9-month monitor time after the EVO ejection. The black dots indicate the corresponding variable concentrations in the control well at the same time points. (*The concentration of U(VI) is measured in μM instead of mM.)

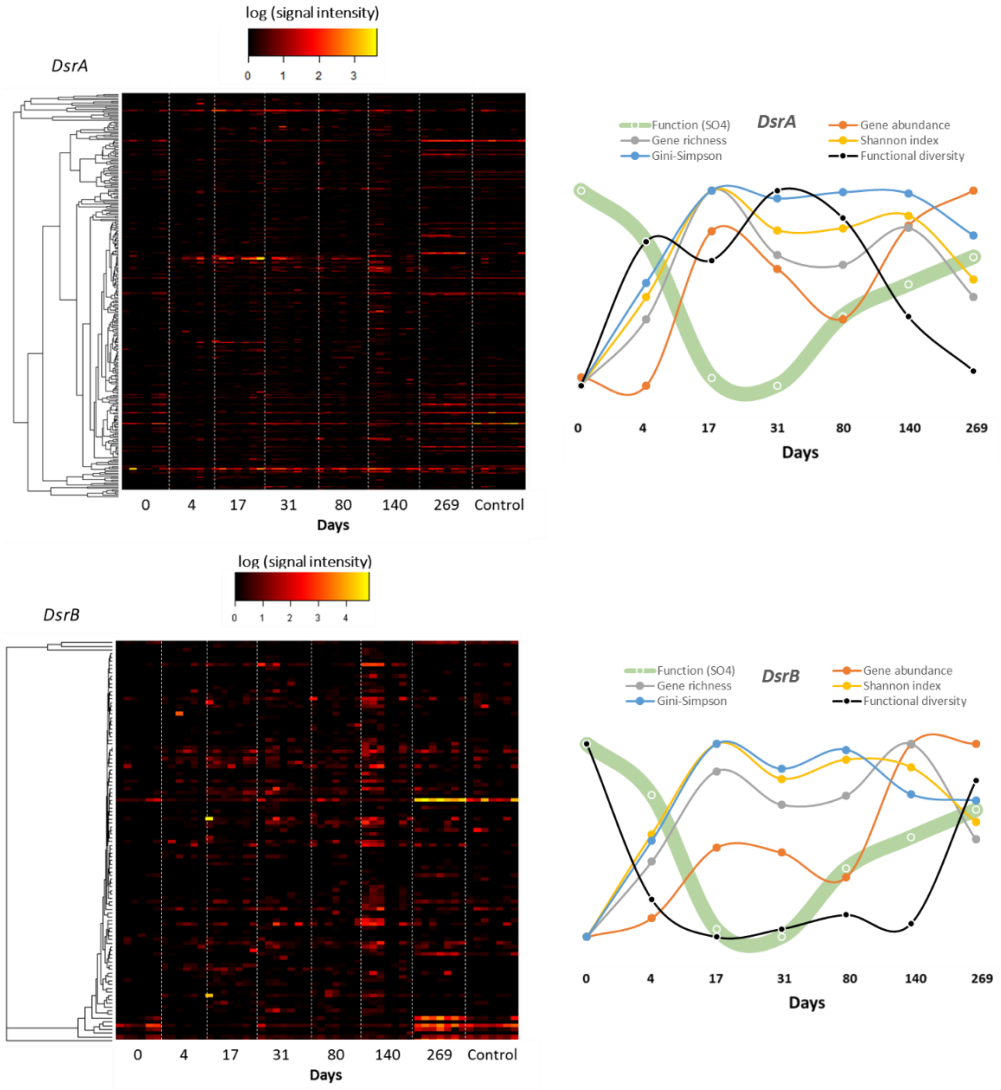


Figure S4. The detailed GeoChip functional profile for *dsrA* (upper) and *dsrB* (lower) genes, and their diversity indices change across time.

Appendix B Supplementary Tables

Table S1. Mock bacterial community species and details

Sequence name	Taxonomy (Phyla or class)	Source	Insert length (nt)[b]
Acidobacteria	Acidobacteria	Drinking water	1359
Actinobacteria	Actinobacteria	Wastewater reactor	1392
Bacteroidetes clone 1	Bacteroidetes	Wastewater reactor	1355
Bacteroidetes clone 2	Bacteroidetes	Drinking water	1352
Caldisericum exile	OP5	DSMZ culture collection-13637	1426
Chlorobi	Chlorobi	Surface water	1374
Cyanobacteria	Cyanobacteria	Surface water	1324
Deferribacter desulfuricans	Deferribacteres	DSMZ culture collection-14783	1410
Deinococcus indicus	Deinococcus-Thermus	DSMZ culture collection-1537	1366
Desulfurispirillum alkaliphilum	Chrysiogenetes	DSMZ culture collection-1827	1375
Dictyoglomus thermophilum	Dictyoglomi	DSMZ culture collection-396	1415
Fibrobacter succinogenes S85	Fibrobacteres	Donated by Isaac Cann, University of Illinois-Urbana Champaign	1372
Gemmatimonadetes	Gemmatimonadetes	Wastewater reactor	1360
Leptotrichia hofstadii	Fusobacteria	DSMZ culture collection-21561	1367
Mycoplasma orale	Firmicutes	DSMZ culture collection-1915	1375
Nitrospira	Nitrospirae	Wastewater reactor	1376
Persephonella hydrogeniphiia H3	Aquificae	Donated by Anne Louise Reysenbach, Portland State University	1389
Planctomycetes	Planctomycetes	Wastewater reactor	1376
Protochlamydia amoebophilia	Chlamydiae	Donated by Mathias Horn, University of Vienna	1360
Spirochaetes	Spirochaetae	Surface water	1396
Sulfurihydrogenibium yellowstonense	Aquificae	Donated by Anne Louise Reysenbach, Portland State University	1378
Synergistetes	Synergistetes	Surface water	1355
Syntrophobacter fumaroxidans	Deltaproteobacteria	Donated by Syed Hashsham, Michigan State University (DSMZ# 117)	1415
Syntrophococcus sucromutans	Firmicutes	Donated by Syed Hashsham, Michigan State University (ATCC# 43584)	1380
Syntrophomonas bryantii	Firmicutes	Donated by Syed Hashsham, Michigan State University (DSMZ# 314A)	1412
Syntrophothermus lipocalidus	Firmicutes	Donated by Syed Hashsham, Michigan State University (DSMZ# 1268)	1500
Syntrophus buswellii	Deltaproteobacteria	Donated by Syed Hashsham, Michigan State University	1413

		(DSMZ# 2612A)	
Syntrophus gentianae	Deltaproteobacteria	Donated by Syed Hashsham, Michigan State University (DMZ# 8423)	1412
Thermodesulfobacterium commune	Thermodesulfobacteria	DSMZ culture collection-2178	1422
Thermomicrobium roseum	Chloroflexi	DSMZ culture collection-5159	1371
Thermotoga neapolitana	Thermotogae.	Donated by Claire Vielle, Michigan State University	1412
Verrucomicrobia	Verrucomicrobia	Surface water	1379
Victivallis vadensis	Lentisphaerae	DSMZ culture collection-8748	1360

[a] The mock community was a gift from Dr. Lutgarde Raskin, Department of Civil and Environmental Engineering, University of Michigan, United States of America.

[b] The insertions are near full length 16S rDNA sequences.

Table S2. Correlations of the strain relative abundance between different sequencing strategies.

Mock community	Sequencing strategy	Correlation (r)*	
		old primer	one-step
Mock1 (Bm1)	**one-step**	0.6858	-
	two-step	0.5949	0.8633
Mock2 (Bm2)	**one-step**	0.9297	-
	two-step	0.9426	0.9746
Mock3 (Bm3)	**one-step**	0.9342	-
	two-step	0.9024	0.9640

(* All the correlations are significant with p-value < 0.001)

Table S3. Non-parametric multivariate dissimilarity tests of fungal microbial community structure across six forest sites and between any two sites.

Sites	MRPP				anosim				adonis			
	Jaccard		Bray Curtis		Jaccard		Bray Curtis		Jaccard		Bray Curtis	
	Delta	p	Delta	p	R	p	R	p	F	p	F	p
Whole	0.780	0.001	0.993	0.001	11.797	0.001	0.838	0.001	0.969	0.001	8.578	0.001
BCI vs LUQ	0.836	0.001	0.995	0.001	6.476	0.001	0.878	0.001	0.884	0.001	5.223	0.001
BCI vs CWT	0.797	0.001	1.000	0.001	11.898	0.001	0.861	0.001	1.000	0.001	7.878	0.001
BCI vs AND	0.787	0.001	1.000	0.001	12.961	0.001	0.850	0.001	0.998	0.001	8.588	0.001
BCI vs HFR	0.789	0.001	1.000	0.001	12.776	0.001	0.829	0.001	1.000	0.001	9.985	0.001
BCI vs NWT	0.767	0.001	1.000	0.001	14.521	0.001	0.818	0.001	0.979	0.001	10.654	0.001
LUQ vs CWT	0.814	0.001	0.988	0.001	9.960	0.001	0.878	0.001	0.983	0.001	6.604	0.001
LUQ vs AND	0.805	0.001	1.000	0.001	11.356	0.001	0.867	0.001	1.000	0.001	7.459	0.001
LUQ vs HFR	0.806	0.001	1.000	0.001	11.083	0.001	0.845	0.001	1.000	0.001	8.704	0.001
LUQ vs NWT	0.784	0.001	1.000	0.001	13.154	0.001	0.834	0.001	1.000	0.001	9.641	0.001
CWT vs AND	0.765	0.001	1.000	0.001	11.990	0.001	0.850	0.001	0.977	0.001	7.202	0.001
CWT vs HFR	0.767	0.001	0.997	0.001	9.508	0.001	0.829	0.001	0.942	0.001	7.850	0.001
CWT vs NWT	0.745	0.001	1.000	0.001	14.992	0.001	0.817	0.001	1.000	0.001	9.926	0.001
AND vs HFR	0.757	0.001	0.999	0.001	11.786	0.001	0.817	0.001	0.972	0.001	9.329	0.001
AND vs NWT	0.735	0.001	0.999	0.001	12.200	0.001	0.806	0.001	0.909	0.001	8.760	0.001
HFR vs NWT	0.737	0.001	1.000	0.001	13.796	0.001	0.785	0.001	0.998	0.001	12.177	0.001

MRPP, multi-response permutation procedures; Adonis, permutational multivariate analysis of variance using distance matrices; ANOSIM, analysis of similarity. Results presented are based on distance matrices calculated with Bray-Curtis and Jaccard index. All tests are significant with p-values < 0.05.

Table S4. The distribution of fungal microbial communities across six forest sites based on their growth morphology

Growth Morphology	AND	BCI	CWT	LUQ	HFR	NWT
- [a]	146305	68115	152085	93479	153239	139365
NULL [a]	39989	41600	65266	49986	67687	50711
Agaricoid	49495	10144	51479	9432	102300	54698
Resupinate	34691	3238	18320	1008	12381	58428
Microfungus	8136	19364	4979	36976	16550	26329
DarkSeptateEndophyte	18799	526	11577	93	10531	20381
Clavarioid	17871	289	12950	335	6863	7780
Gasteroid	14755	7101	6167	903	3570	11258
FacultativeYeast	7056	2578	1145	6032	1238	7479
Agaricoid,Corticioid,Gasteroid, Pleurotoid,orSecotioid	6210	21	11104	578	5763	15
CupFungus	9805	10	734	7	351	9686
Hydnoid	1504	3856	7876	3	160	3671
Boletoid	3215	132	756	10	8612	2325
Yeast	1477	1705	1392	3024	1127	414
Tremelloid-Yeast	816	393	1732	3409	660	256
Agaricoid-Polyporoid	90	2	2	2	9	6777
Agaricoid-Gasteroid-Secotioid	165	4860	44	77	2	1
DarkSeptateMicrofungus	50	0	850	1	2351	7
Polyporoid	26	437	62	377	21	0
Thallus	24	32	58	6	17	393
FacultativeYeast-Microfungus	1	0	92	0	36	18
CorticioidFungusorThallus	18	1	0	0	103	16
Phalloid	0	0	0	31	0	0
Microfungus;FacultativeYeast (Tedersooetal.2014)	4	0	5	0	2	7
Corticioid	0	0	0	0	8	0
No species information [b]	53765	249863	65592	208498	20686	14252

[a] The '-' and 'NULL' annotations from FunGuild program indicates growth morphology information was not included in the database.

[b] For the sequences cannot be classified into species (no hit in the references), the functional and growth morphology information cannot be predicted.

Table S5. Fungal richness predictors in final multiple linear regression models

Variables	Coefficients	*Pr*(>|t|) [a]	R^2	AdjR^2Cum	p-value[b]
Lat	-0.401	***	0.595	0.5919	0.001
pH	0.313	***	0.056	0.6458	0.001
Total nitrogen	-0.296	***	0.018	0.6618	0.012
Total carbon	-0.259	**	0.014	0.6739	0.018
Temp	0.855	***	0.010	0.6821	0.042
Precipitation	-0.623	***	0.065	0.7476	0.001

[a]: significant level of multiple linear regression
[b]: forward selection α criteria, must <0.05 to ensure the significance of the model

Table S6. Significant tests (PERMAONOVA*) of the overall community functional structure changes before and after EVO injection

Community groups	All probes		FD probes		Gene richness		Gene abundance		Shannon index (H')		Functional diversity	
	F	p	F	p	F	p	F	p	F	p	F	p
All	3.862	**0.001**	3.874	**0.001**	3.755	**0.001**	3.305	**0.001**	4.029	**0.001**	3.754	**0.001**
Day0 vs Day4	3.408	**0.001**	3.423	**0.009**	1.964	**0.020**	1.812	**0.014**	2.248	**0.028**	2.328	**0.011**
Day0 vs Day17	5.346	**0.002**	5.365	**0.002**	3.777	**0.001**	4.138	**0.001**	5.565	**0.002**	5.903	**0.001**
Day0 vs Day31	2.496	**0.021**	2.519	**0.015**	3.088	**0.004**	2.942	**0.009**	3.506	**0.005**	3.684	**0.002**
Day0 vs Day80	2.959	**0.007**	2.985	**0.001**	3.662	**0.003**	3.739	**0.002**	4.415	**0.001**	4.678	**0.002**
Day0 vs Day140	3.325	**0.010**	3.340	**0.010**	3.109	**0.004**	3.609	**0.004**	3.525	**0.006**	3.608	**0.010**
Day0 vs Day269	3.551	**0.014**	3.547	**0.009**	1.937	**0.033**	1.834	**0.029**	2.076	**0.028**	2.112	**0.025**
Day0 vs Ctrl	3.437	**0.019**	3.417	**0.016**	2.198	**0.011**	2.150	**0.018**	2.167	**0.022**	2.243	**0.014**
Day4 vs Day17	2.136	**0.009**	2.122	**0.011**	1.560	**0.015**	1.375	**0.020**	1.944	**0.029**	1.923	**0.033**
Day4 vs Day31	2.017	**0.014**	2.005	**0.013**	1.581	**0.020**	1.332	**0.035**	1.323	0.138	1.322	0.151
Day4 vs Day80	1.852	**0.004**	1.836	**0.003**	1.798	**0.001**	1.506	**0.002**	1.807	**0.008**	1.796	**0.013**
Day4 vs Day140	2.178	**0.023**	2.204	**0.011**	1.286	0.135	1.277	0.081	1.553	0.096	1.523	0.118
Day4 vs Day269	6.799	**0.002**	6.791	**0.001**	1.937	**0.002**	1.772	**0.001**	2.316	**0.003**	2.269	**0.002**
Day4 vs Ctrl	7.576	**0.001**	7.523	**0.002**	1.954	**0.009**	1.637	**0.009**	1.984	**0.035**	1.980	**0.040**
Day17 vs Day31	2.345	**0.006**	2.350	**0.006**	1.232	0.187	1.403	0.098	1.256	0.218	1.251	0.195
Day17 vs Day80	2.522	**0.001**	2.527	**0.003**	1.032	0.502	1.086	0.432	2.134	**0.019**	2.095	**0.017**
Day17 vs Day140	3.084	**0.001**	3.118	**0.001**	1.040	0.453	1.221	0.280	2.026	**0.034**	1.952	**0.031**
Day17 vs Day269	9.030	**0.003**	9.070	**0.001**	2.661	**0.003**	3.142	**0.004**	4.870	**0.002**	4.652	**0.002**
Day17 vs Ctrl	10.247	**0.002**	10.219	**0.002**	2.641	**0.013**	2.691	**0.018**	3.982	**0.004**	3.929	**0.006**
Day31 vs Day80	1.479	0.062	1.498	0.072	0.873	0.733	0.750	0.850	1.133	0.284	1.131	0.291
Day31 vs Day140	1.805	0.061	1.814	**0.049**	0.954	0.503	1.063	0.386	0.984	0.418	0.985	0.428
Day31 vs Day269	4.779	**0.001**	4.806	**0.003**	1.889	**0.011**	2.118	**0.008**	2.591	**0.006**	2.539	**0.006**
Day31 vs Ctrl	4.983	**0.001**	4.969	**0.002**	1.836	**0.041**	1.702	0.064	1.876	0.060	1.876	0.069
Day80 vs Day140	1.476	0.121	1.498	0.098	1.163	0.239	1.255	0.259	1.125	0.301	1.117	0.297
Day80 vs Day269	5.447	**0.001**	5.483	**0.002**	2.516	**0.001**	2.524	**0.001**	3.546	**0.002**	3.427	**0.001**
Day80 vs Ctrl	6.843	**0.002**	6.870	**0.001**	2.794	**0.003**	2.453	**0.020**	3.086	**0.004**	3.068	**0.002**
Day140 vs Day269	5.733	**0.001**	5.777	**0.003**	1.985	**0.028**	2.576	**0.009**	2.648	**0.012**	2.528	**0.009**
Day140 vs Ctrl	7.017	**0.002**	7.025	**0.001**	2.327	**0.023**	2.773	**0.018**	2.175	**0.045**	2.121	0.051
Day269 vs Ctrl	3.838	**0.002**	3.855	**0.001**	0.832	0.690	0.881	0.590	1.078	0.360	1.076	0.320

*Permutational multivariate analysis of variance using distance matrices uses 'adonis' function from 'vegan' package in R. Significance tests were carried out using *F*-tests based on sequential sums of squares from permutations of the raw data. Group 'Ctrl' are the samples collected from well 8 across different time points. All the p-values < 0.05 are marked in bold.

Table S7. Multiple regression on distance matrices analysis of microbial community functional diversity and environmental variables

	All genes		Carbon cycling		Nitrogen cycling		Energy process		Metal Resistance		Organic Remediation		Phosphorus		Sulphur		Antibiotic resistance		Other categories	
	R^2	p	R^2	p	R^2	p	R^2	p	R^2	p	R^2	p	R^2	p	R^2	p	R^2	p	R^2	p
Whole model	0.140	**0.003**	0.129	**0.026**	0.100	0.291	0.179	**0.040**	0.101	0.080	0.138	**0.018**	0.126	**0.001**	0.118	**0.007**	0.114	0.102	0.116	0.216
pH	0.000	0.868	0.000	0.908	0.000	0.810	0.002	0.581	0.003	0.335	0.003	0.381	0.005	0.157	0.001	0.497	0.003	0.435	0.003	0.434
Acetate	0.019	**0.026**	0.004	0.349	0.003	0.484	0.002	0.571	0.025	**0.015**	0.018	0.061	0.021	**0.012**	0.003	0.298	0.001	0.614	0.009	0.219
NO_3^-	0.000	0.758	0.006	0.235	0.005	0.367	0.000	0.917	0.010	0.108	0.005	0.327	0.028	**0.004**	0.021	**0.016**	0.007	0.243	0.016	0.087
SO_4^{2-}	0.033	**0.001**	0.008	**0.012**	0.022	**0.001**	0.021	**0.002**	0.041	**0.001**	0.015	**0.003**	0.038	**0.001**	0.058	**0.001**	0.017	**0.002**	0.003	0.157
Fe(II)	0.020	**0.033**	0.014	0.121	0.004	0.426	0.082	**0.004**	0.004	0.359	0.023	**0.037**	0.005	0.215	0.002	0.517	0.000	0.875	0.004	0.452
U(VI)	0.079	**0.001**	0.040	**0.002**	0.041	**0.001**	0.009	0.089	0.060	**0.002**	0.058	**0.001**	0.012	**0.008**	0.026	**0.002**	0.013	**0.018**	0.005	0.150
Cl	0.000	0.880	0.003	0.353	0.000	0.804	0.039	**0.020**	0.000	0.812	0.000	0.898	0.000	0.992	0.001	0.625	0.000	0.878	0.015	0.089
Ag	0.001	0.471	0.001	0.480	0.014	**0.004**	0.002	0.231	0.002	0.139	0.000	0.954	0.001	0.205	0.001	0.282	0.001	0.526	0.000	0.542
Al	0.011	**0.016**	0.016	**0.012**	0.010	**0.040**	0.003	0.275	0.002	0.397	0.006	0.096	0.001	0.356	0.000	0.718	0.008	0.052	0.002	0.423
Ba	0.006	0.140	0.000	0.997	0.000	0.849	0.011	0.104	0.013	**0.028**	0.007	0.147	0.012	**0.011**	0.002	0.270	0.001	0.625	0.000	0.940
Ca	0.002	0.509	0.006	0.220	0.005	0.347	0.014	0.120	0.003	0.381	0.000	0.811	0.004	0.246	0.001	0.690	0.008	0.155	0.004	0.418
Cr	0.003	0.395	0.001	0.583	0.003	0.450	0.009	0.211	0.001	0.595	0.000	0.803	0.002	0.300	0.002	0.442	0.032	**0.004**	0.000	0.950
Ga	0.017	**0.011**	0.001	0.571	0.001	0.574	0.030	**0.010**	0.016	**0.012**	0.020	**0.018**	0.031	**0.001**	0.011	**0.019**	0.007	0.100	0.000	0.870
K	0.000	0.867	0.004	0.353	0.001	0.723	0.004	0.482	0.000	0.740	0.000	0.820	0.000	0.909	0.000	0.880	0.010	0.197	0.001	0.707
Mg	0.000	0.825	0.002	0.513	0.003	0.523	0.000	0.869	0.001	0.720	0.002	0.614	0.003	0.342	0.002	0.476	0.005	0.369	0.000	0.823
Mn(II)	0.005	0.288	0.001	0.688	0.002	0.580	0.046	**0.032**	0.005	0.354	0.003	0.480	0.007	0.157	0.002	0.548	0.002	0.577	0.000	0.999
Sr	0.000	0.888	0.001	0.730	0.000	0.858	0.003	0.550	0.011	0.126	0.001	0.764	0.001	0.618	0.000	0.942	0.001	0.611	0.002	0.629
Zn	0.008	0.257	0.023	0.075	0.014	0.159	0.001	0.818	0.015	0.099	0.000	0.925	0.000	0.961	0.005	0.225	0.000	0.848	0.010	0.142

Table S8 Mantel test of correlation between differences in detailed microbial functional structures and environmental variables (p<0.05 is marked in bold font)

Function categories	Gene Abundance		Gene richness		Shannon index		Gini-Simpson		Functional diversity		FD probes	
	rho	p	rho	p	rho	p	rho	p	rho	p	rho	p
Antibiotic resistance	0.008	0.416	0.010	0.416	0.033	0.305	0.070	0.172	0.098	0.085	0.010	0.433
Carbon cycling	-0.013	0.541	0.073	0.143	0.069	0.151	0.062	0.184	0.093	0.086	-0.014	0.563
Energy process	0.003	0.456	0.068	0.145	0.053	0.211	0.035	0.310	0.135	**0.041**	-0.052	0.769
Metal Resistance	-0.010	0.533	0.067	0.165	0.096	0.096	0.134	**0.043**	0.156	**0.022**	0.018	0.395
Nitrogen	-0.001	0.474	0.066	0.172	0.083	0.141	0.113	0.088	0.140	**0.045**	0.020	0.386
Organic Remediation	-0.010	0.529	0.056	0.216	0.066	0.170	0.063	0.187	0.065	0.173	0.012	0.422
Phosphorus	0.018	0.347	0.073	0.149	0.108	0.080	0.098	0.086	0.118	**0.007**	0.040	0.282
Sulphur	-0.029	0.650	0.057	0.205	0.025	0.352	-0.002	0.481	0.081	0.067	0.019	0.391
Other category	-0.009	0.523	0.052	0.187	0.038	0.265	-0.029	0.623	-0.071	0.846	0.105	0.058

References

Ahn, S. J., J. Costa and J. R. Emanuel (1996). "PicoGreen quantitation of DNA: effective evaluation of samples pre- or post-PCR." Nucleic Acids Res **24**(13): 2623-2625.

Altschul, S. F., W. Gish, W. Miller, E. W. Myers and D. J. Lipman (1990). "Basic Local Alignment Search Tool." Journal of Molecular Biology **215**(3): 403-410.

Balvanera, P., A. B. Pfisterer, N. Buchmann, J. S. He, T. Nakashizuka, D. Raffaelli and B. Schmid (2006). "Quantifying the evidence for biodiversity effects on ecosystem functioning and services." Ecology Letters **9**(10): 1146-1156.

Bartram, A. K., M. D. J. Lynch, J. C. Stearns, G. Moreno-Hagelsieb and J. D. Neufeld (2011). "Generation of Multimillion-Sequence 16S rRNA Gene Libraries from Complex Microbial Communities by Assembling Paired-End Illumina Reads (vol 77, pg 3846, 2011)." Applied and Environmental Microbiology **77**(15): 5569-5569.

Bellemain, E., T. Carlsen, C. Brochmann, E. Coissac, P. Taberlet and H. Kauserud (2010). "ITS as an environmental DNA barcode for fungi: an in silico approach reveals potential PCR biases." Bmc Microbiology **10**.

Bellwood, D. R., A. S. Hoey and J. H. Choat (2003). "Limited functional redundancy in high diversity systems: resilience and ecosystem function on coral reefs." Ecology Letters **6**(4): 281-285.

Belotte, D., J. B. Curien, R. C. Maclean and G. Bell (2003). "An experimental test of local adaptation in soil bacteria." Evolution **57**(1): 27-36.

Bennett, E. M., W. Cramer, A. Begossi, G. Cundill, S. Diaz, B. N. Egoh, I. R. Geijzendorffer, C. B. Krug, S. Lavorel, E. Lazos, L. Lebel, B. Martin-Lopez, P. Meyfroidt, H. A. Mooney, J. L. Nel, U. Pascual, K. Payet, N. P. Harguindeguy, G. D. Peterson, A. H. N. Prieur-Richard, B. Reyers, P. Roebeling, R. Seppelt, M. Solan, P. Tschakert, T. Tscharntke, B. L. Turner, P. H. Verburg, E. F. Viglizzo, P. C. L. White and G. Woodward (2015). "Linking biodiversity, ecosystem services, and human well-being: three challenges for designing research for sustainability." Current Opinion in Environmental Sustainability **14**: 76-85.

Blankenberg, D., A. Gordon, G. Von Kuster, N. Coraor, J. Taylor, A. Nekrutenko and T. Galaxy (2010). "Manipulation of FASTQ data with Galaxy." Bioinformatics **26**(14): 1783-1785.

Blondel, J. (2003). "Guilds or functional groups: does it matter?" Oikos **100**(2): 223-231.

Botta-Dukat, Z. (2005). "Rao's quadratic entropy as a measure of functional diversity based on multiple traits." Journal of Vegetation Science **16**(5): 533-540.

Bouchez, T., A. L. Blieux, S. Dequiedt, I. Domaizon, A. Dufresne, S. Ferreira, J. J. Godon, J. Hellal, C. Joulian, A. Quaiser, F. Martin-Laurent, A. Mauffret, J. M. Monier, P. Peyret, P. Schmitt-Koplin, O. Sibourg, E. D'oiron, A. Bispo, I. Deportes, C. Grand, P. Cuny, P. A. Maron and L. Ranjard (2016). "Molecular microbiology methods for environmental diagnosis." Environmental Chemistry Letters **14**(4): 423-441.

Bracken, M. E. S., S. E. Friberg, C. A. Gonzalez-Dorantes and S. L. Williams (2008). "Functional consequences of realistic biodiversity changes in a marine ecosystem." Proceedings of the National Academy of Sciences of the United States of America **105**(3): 924-928.

Buee, M., M. Reich, C. Murat, E. Morin, R. H. Nilsson, S. Uroz and F. Martin (2009). "454 Pyrosequencing analyses of forest soils reveal an unexpectedly high fungal diversity." New Phytologist **184**(2): 449-456.

Caporaso, J. G., J. Kuczynski, J. Stombaugh, K. Bittinger, F. D. Bushman, E. K. Costello, N. Fierer, A. G. Pena, J. K. Goodrich, J. I. Gordon, G. A. Huttley, S. T. Kelley, D. Knights, J. E. Koenig, R. E. Ley, C. A. Lozupone, D. McDonald, B. D. Muegge, M. Pirrung, J. Reeder, J. R. Sevinsky, P. J. Turnbaugh, W. A. Walters, J. Widmann, T. Yatsunenko, J. Zaneveld and R. Knight (2010). "QIIME allows analysis of high-throughput community sequencing data." Nat Methods **7**(5): 335-336.

Caporaso, J. G., C. L. Lauber, W. A. Walters, D. Berg-Lyons, J. Huntley, N. Fierer, S. M. Owens, J. Betley, L. Fraser, M. Bauer, N. Gormley, J. A. Gilbert, G. Smith and R. Knight (2012). "Ultra-high-throughput microbial community analysis on the Illumina HiSeq and MiSeq platforms." Isme Journal **6**(8): 1621-1624.

Chourey, K., S. Nissen, T. Vishnivetskaya, M. Shah, S. Pfiffner, R. L. Hettich and F. E. Loffler (2013). "Environmental proteomics reveals early microbial community responses to biostimulation at a uranium- and nitrate-contaminated site." Proteomics **13**(18-19): 2921-2930.

Clark, C. M., D. F. B. Flynn, B. J. Butterfield and P. B. Reich (2012). "Testing the Link between Functional Diversity and Ecosystem Functioning in a Minnesota Grassland Experiment." Plos One **7**(12).

Colin, N., S. Villeger, M. Wilkes, A. de Sostoa and A. Maceda-Veiga (2018). "Functional diversity measures revealed impacts of non-native species and habitat degradation on species-poor freshwater fish assemblages." Science of the Total Environment **625**: 861-871.

Colwell, R. R. (1997). "Microbial diversity: The importance of exploration and conservation." Journal of Industrial Microbiology & Biotechnology **18**(5): 302-307.

Cornelissen, J. H. C., S. Lavorel, E. Garnier, S. Diaz, N. Buchmann, D. E. Gurvich, P. B. Reich, H. ter Steege, H. D. Morgan, M. G. A. van der Heijden, J. G. Pausas and H. Poorter (2003). "A handbook of protocols for standardised and easy measurement of plant functional traits worldwide." Australian Journal of Botany **51**(4): 335-380.

Cornwell, W. K., D. W. Schwilk and D. D. Ackerly (2006). "A trait-based test for habitat filtering: Convex hull volume." Ecology **87**(6): 1465-1471.

Dannemiller, K. C., D. Reeves, K. Bibby, N. Yamamoto and J. Peccia (2014). "Fungal High-throughput Taxonomic Identification tool for use with Next-Generation Sequencing (FHiTINGS)." Journal of Basic Microbiology **54**(4): 315-321.

de Bello, F., S. Lavergne, C. N. Meynard, J. Leps and W. Thuiller (2010). "The partitioning of diversity: showing Theseus a way out of the labyrinth." Journal of Vegetation Science **21**(5): 992-1000.

de Bello, F., J. Leps, S. Lavorel and M. Moretti (2007). "Importance of species abundance for assessment of trait composition: an example based on pollinator communities." Community Ecology **8**(2): 163-170.

de Vienne, D. M., G. Aguileta and S. Ollier (2011). "Euclidean Nature of Phylogenetic Distance Matrices." Systematic Biology **60**(6): 826-832.

Delgado-Baquerizo, M., F. T. Maestre, P. B. Reich, T. C. Jeffries, J. J. Gaitan, D. Encinar, M. Berdugo, C. D. Campbell and B. K. Singh (2016). "Microbial diversity drives multifunctionality in terrestrial ecosystems." Nature Communications **7**.

DeSantis, T. Z., P. Hugenholtz, N. Larsen, M. Rojas, E. L. Brodie, K. Keller, T. Huber, D. Dalevi, P. Hu and G. L. Andersen (2006). "Greengenes, a chimera-checked 16S rRNA gene database and workbench compatible with ARB." Appl Environ Microbiol **72**(7): 5069-5072.

Dethlefsen, L., S. Huse, M. L. Sogin and D. A. Relman (2008). "The Pervasive Effects of an Antibiotic on the Human Gut Microbiota, as Revealed by Deep 16S rRNA Sequencing." Plos Biology **6**(11): 2383-2400.

Edgar, R. C. (2010). "Search and clustering orders of magnitude faster than BLAST." Bioinformatics **26**(19): 2460-2461.

Edgar, R. C. (2010). "Search and clustering orders of magnitude faster than BLAST." Bioinformatics **26**(19): 2460-2461.

Edgar, R. C. (2013). "UPARSE: highly accurate OTU sequences from microbial amplicon reads." Nat Methods **10**(10): 996-998.

Edgar, R. C., B. J. Haas, J. C. Clemente, C. Quince and R. Knight (2011). "UCHIME improves sensitivity and speed of chimera detection." Bioinformatics **27**(16): 2194-2200.

Fadrosh, D. W., B. Ma, P. Gajer, N. Sengamalay, S. Ott, R. M. Brotman and J. Ravel (2014). "An improved dual-indexing approach for multiplexed 16S rRNA gene sequencing on the Illumina MiSeq platform." Microbiome **2**(1): 6.

Fadrosh, D. W., B. Ma, P. Gajer, N. Sengamalay, S. Ott, R. M. Brotman and J. Ravel (2014). "An improved dual-indexing approach for multiplexed 16S rRNA gene sequencing on the Illumina MiSeq platform." Microbiome **2**.

Faith, J. J., J. L. Guruge, M. Charbonneau, S. Subramanian, H. Seedorf, A. L. Goodman, J. C. Clemente, R. Knight, A. C. Heath, R. L. Leibel, M. Rosenbaum and J. I. Gordon (2013). "The Long-Term Stability of the Human Gut Microbiota." Science **341**(6141): 44-+.

Feinstein, L. M. and C. B. Blackwood (2012). "Taxa-area relationship and neutral dynamics influence the diversity of fungal communities on senesced tree leaves." Environ Microbiol **14**(6): 1488-1499.

Fetzer, I., K. Johst, R. Schawe, T. Banitz, H. Harms and A. Chatzinotas (2015). "The extent of functional redundancy changes as species' roles shift in different environments." Proceedings of the National Academy of Sciences of the United States of America **112**(48): 14888-14893.

Fierer, N., M. S. Strickland, D. Liptzin, M. A. Bradford and C. C. Cleveland (2009). "Global patterns in belowground communities." Ecol Lett **12**(11): 1238-1249.

Finlay, B. J. (2002). "Global dispersal of free-living microbial eukaryote species." Science **296**(5570): 1061-1063.

Fitter, A. H., C. A. Gilligan, K. Hollingworth, A. Kleczkowski, R. M. Twyman, J. W. Pitchford and N. S. B. Programme (2005). "Biodiversity and ecosystem function in soil." Functional Ecology **19**(3): 369-377.

Fu, L., B. Niu, Z. Zhu, S. Wu and W. Li (2012). "CD-HIT: accelerated for clustering the next-generation sequencing data." Bioinformatics **28**(23): 3150-3152.

Gaby, J. C., L. Rishishwar, L. C. Valderrama-Aguirre, S. J. Green, A. Valderrama-Aguirre, I. K. Jordan and J. E. Kostka (2018). "Diazotroph Community Characterization via a High-Throughput nifH Amplicon Sequencing and Analysis Pipeline." Appl Environ Microbiol **84**(4).

Gagic, V., I. Bartomeus, T. Jonsson, A. Taylor, C. Winqvist, C. Fischer, E. M. Slade, I. Steffan-Dewenter, M. Emmerson, S. G. Potts, T. Tscharntke, W. Weisser and R. Bommarco (2015). "Functional identity and diversity of animals predict ecosystem functioning better than species-based indices." Proceedings of the Royal Society B-Biological Sciences **282**(1801).

Gibson, J., S. Shokralla, T. M. Porter, I. King, S. van Konynenburg, D. H. Janzen, W. Hallwachs and M. Hajibabaei (2014). "Simultaneous assessment of the macrobiome and

microbiome in a bulk sample of tropical arthropods through DNA metasystematics." Proceedings of the National Academy of Sciences of the United States of America **111**(22): 8007-8012.

Gihring, T. M., G. X. Zhang, C. C. Brandt, S. C. Brooks, J. H. Campbell, S. Carroll, C. S. Criddle, S. J. Green, P. Jardine, J. E. Kostka, K. Lowe, T. L. Mehlhorn, W. Overholt, D. B. Watson, Z. M. Yang, W. M. Wu and C. W. Schadt (2011). "A Limited Microbial Consortium Is Responsible for Extended Bioreduction of Uranium in a Contaminated Aquifer." Applied and Environmental Microbiology **77**(17): 5955-5965.

Goodwin, S., J. D. McPherson and W. R. McCombie (2016). "Coming of age: ten years of next-generation sequencing technologies." Nature Reviews Genetics **17**(6): 333-351.

Graham, E. B., J. E. Knelman, A. Schindlbacher, S. Siciliano, M. Breulmann, A. Yannarell, J. M. Bemans, G. Abell, L. Philippot, J. Prosser, A. Foulquier, J. C. Yuste, H. C. Glanville, D. L. Jones, F. Angel, J. Salminen, R. J. Newton, H. Burgmann, L. J. Ingram, U. Hamer, H. M. P. Siljanen, K. Peltoniemi, K. Potthast, L. Baneras, M. Hartmann, S. Banerjee, R. Q. Yu, G. Nogaro, A. Richter, M. Koranda, S. C. Castle, M. Goberna, B. Song, A. Chatterjee, O. C. Nunes, A. R. Lopes, Y. P. Cao, A. Kaisermann, S. Hallin, M. S. Strickland, J. Garcia-Pausas, J. Barba, H. Kang, K. Isobe, S. Papaspyrou, R. Pastorelli, A. Lagomarsino, E. S. Lindstrom, N. Basiliko and D. R. Nemergut (2016). "Microbes as Engines of Ecosystem Function: When Does Community Structure Enhance Predictions of Ecosystem Processes?" Frontiers in Microbiology **7**.

Green, J. L., A. J. Holmes, M. Westoby, I. Oliver, D. Briscoe, M. Dangerfield, M. Gillings and A. J. Beattie (2004). "Spatial scaling of microbial eukaryote diversity." Nature **432**(7018): 747-750.

Grice, E. A., H. H. Kong, S. Conlan, C. B. Deming, J. Davis, A. C. Young, G. G. Bouffard, R. W. Blakesley, P. R. Murray, E. D. Green, M. L. Turner, J. A. Segre and N. C. S. Progra (2009). "Topographical and Temporal Diversity of the Human Skin Microbiome." Science **324**(5931): 1190-1192.

Grum-Grzhimaylo, A. A., M. L. Georgieva, S. A. Bondarenko, A. J. M. Debets and E. N. Bilanenko (2016). "On the diversity of fungi from soda soils." Fungal Diversity **76**(1): 27-74.

Hardy, O. J. and B. Senterre (2007). "Characterizing the phylogenetic structure of communities by an additive partitioning of phylogenetic diversity." Journal of Ecology **95**(3): 493-506.

He, Z. L., Y. Deng, J. D. Van Nostrand, Q. C. Tu, M. Y. Xu, C. L. Hemme, X. Y. Li, L. Y. Wu, T. J. Gentry, Y. F. Yin, J. Liebich, T. C. Hazen and J. Z. Zhou (2010). "GeoChip 3.0 as a high-throughput tool for analyzing microbial community composition, structure and functional activity." Isme Journal **4**(9): 1167-1179.

Heemsbergen, D. A., M. P. Berg, M. Loreau, J. R. van Haj, J. H. Faber and H. A. Verhoef (2004). "Biodiversity effects on soil processes explained by interspecific functional dissimilarity." Science **306**(5698): 1019-1020.

Herlemann, D. P. R., M. Labrenz, K. Jurgens, S. Bertilsson, J. J. Waniek and A. F. Andersson (2011). "Transitions in bacterial communities along the 2000 km salinity gradient of the Baltic Sea." Isme Journal **5**(10): 1571-1579.

Hillebrand, H. (2004). "On the generality of the latitudinal diversity gradient." Am Nat **163**(2): 192-211.

Hogberg, M. N., E. Baath, A. Nordgren, K. Arnebrant and P. Hogberg (2003). "Contrasting effects of nitrogen availability on plant carbon supply to mycorrhizal fungi and saprotrophs - a hypothesis based on field observations in boreal forest." New Phytologist **160**(1): 225-238.

Hooper, D. U., F. S. Chapin, J. J. Ewel, A. Hector, P. Inchausti, S. Lavorel, J. H. Lawton, D. M. Lodge, M. Loreau, S. Naeem, B. Schmid, H. Setala, A. J. Symstad, J. Vandermeer and D. A. Wardle (2005). "Effects of biodiversity on ecosystem functioning: A consensus of current knowledge." Ecological Monographs **75**(1): 3-35.

Hummelen, R., A. D. Fernandes, J. M. Macklaim, R. J. Dickson, J. Changalucha, G. B. Gloor and G. Reid (2010). "Deep Sequencing of the Vaginal Microbiota of Women with HIV." Plos One **5**(8).

Hummelen, R., J. M. Macklaim, J. E. Bisanz, J. A. Hammond, A. McMillan, R. Vongsa, D. Koenig, G. B. Gloor and G. Reid (2011). "Vaginal microbiome and epithelial gene array in post-menopausal women with moderate to severe dryness." PLoS One **6**(11): e26602.

Jax, K. (2005). "Function and "functioning" in ecology: what does it mean?" Oikos **111**(3): 641-648.

Jost, L. (2007). "Partitioning diversity into independent alpha and beta components." Ecology **88**(10): 2427-2439.

Jukes TH, C. C. (1969). "Evolution of protein molecules." In Munro HN, editor, Mammalian Protein Metabolism: 21-132.

Kandeler, E., C. Kampichler and O. Horak (1996). "Influence of heavy metals on the functional diversity of soil microbial communities." Biology and Fertility of Soils **23**(3): 299-306.

Karimi, B., P. A. Maron, N. C. P. Boure, N. Bernard, D. Gilbert and L. Ranjard (2017). "Microbial diversity and ecological networks as indicators of environmental quality." Environmental Chemistry Letters **15**(2): 265-281.

Koljalg, U., R. H. Nilsson, K. Abarenkov, L. Tedersoo, A. F. S. Taylor, M. Bahram, S. T. Bates, T. D. Bruns, J. Bengtsson-Palme, T. M. Callaghan, B. Douglas, T. Drenkhan, U. Eberhardt, M. Duenas, T. Grebenc, G. W. Griffith, M. Hartmann, P. M. Kirk, P. Kohout, E. Larsson, B. D. Lindahl, R. Luecking, M. P. Martin, P. B. Matheny, N. H. Nguyen, T. Niskanen, J. Oja, K. G. Peay, U. Peintner, M. Peterson, K. Poldmaa, L. Saag, I. Saar, A. Schuessler, J. A. Scott, C. Senes, M. E. Smith, A. Suija, D. L. Taylor, M. T. Telleria, M. Weiss and K. H. Larsson (2013). "Towards a unified paradigm for sequence-based identification of fungi." Molecular Ecology **22**(21): 5271-5277.

Kong, Y. (2011). "Btrim: a fast, lightweight adapter and quality trimming program for next-generation sequencing technologies." Genomics **98**(2): 152-153.

Kozich, J. J., S. L. Westcott, N. T. Baxter, S. K. Highlander and P. D. Schloss (2013). "Development of a Dual-Index Sequencing Strategy and Curation Pipeline for Analyzing Amplicon Sequence Data on the MiSeq Illumina Sequencing Platform." Applied and Environmental Microbiology **79**(17): 5112-5120.

Kozich, J. J., S. L. Westcott, N. T. Baxter, S. K. Highlander and P. D. Schloss (2013). "Development of a dual-index sequencing strategy and curation pipeline for analyzing amplicon sequence data on the MiSeq Illumina sequencing platform." Appl Environ Microbiol **79**(17): 5112-5120.

Krueger, F., S. R. Andrews and C. S. Osborne (2011). "Large Scale Loss of Data in Low-Diversity Illumina Sequencing Libraries Can Be Recovered by Deferred Cluster Calling." Plos One **6**(1).

Lande, R. (1996). "Statistics and partitioning of species diversity, and similarity among multiple communities." Oikos **76**(1): 5-13.

Lauber, C. L., M. Hamady, R. Knight and N. Fierer (2009). "Pyrosequencing-Based Assessment of Soil pH as a Predictor of Soil Bacterial Community Structure at the Continental Scale." Applied and Environmental Microbiology **75**(15): 5111-5120.

Laureto, L. M. O., M. V. Cianciaruso and D. S. M. Samia (2015). "Functional diversity: an overview of its history and applicability." Natureza & Conservacao **13**(2): 112-116.

Legendre, P. and M. J. Anderson (1999). "Distance-based redundancy analysis: Testing multispecies responses in multifactorial ecological experiments (vol 69, pg 1, 1999)." Ecological Monographs **69**(4): 512-512.

Logue, J. B. and E. S. Lindstrom (2010). "Species sorting affects bacterioplankton community composition as determined by 16S rDNA and 16S rRNA fingerprints." ISME J **4**(6): 729-738.

Lozupone, C. A., J. I. Stombaugh, J. I. Gordon, J. K. Jansson and R. Knight (2012). "Diversity, stability and resilience of the human gut microbiota." Nature **489**(7415): 220-230.

Lundberg, D. S., S. Yourstone, P. Mieczkowski, C. D. Jones and J. L. Dangl (2013). "Practical innovations for high-throughput amplicon sequencing." Nature Methods **10**(10): 999-+.

Lundberg, D. S., S. Yourstone, P. Mieczkowski, C. D. Jones and J. L. Dangl (2013). "Practical innovations for high-throughput amplicon sequencing." Nat Methods **10**(10): 999-1002.

Magoc, T. and S. L. Salzberg (2011). "FLASH: fast length adjustment of short reads to improve genome assemblies." Bioinformatics **27**(21): 2957-2963.

Margesin, R., M. Jud, D. Tscherko and F. Schinner (2009). "Microbial communities and activities in alpine and subalpine soils." Fems Microbiology Ecology **67**(2): 208-218.

Martiny, A. C., K. Treseder and G. Pusch (2013). "Phylogenetic conservatism of functional traits in microorganisms." Isme Journal **7**(4): 830-838.

Martiny, J. B. H., J. A. Eisen, K. Penn, S. D. Allison and M. C. Horner-Devine (2011). "Drivers of bacterial beta-diversity depend on spatial scale." Proceedings of the National Academy of Sciences of the United States of America **108**(19): 7850-7854.

Martiny, J. B. H., S. E. Jones, J. T. Lennon and A. C. Martiny (2015). "Microbiomes in light of traits: A phylogenetic perspective." Science **350**(6261).

Mason, N. W. H., K. MacGillivray, J. B. Steel and J. B. Wilson (2003). "An index of functional diversity." Journal of Vegetation Science **14**(4): 571-578.

McCann, K. S. (2000). "The diversity-stability debate." Nature **405**(6783): 228-233.

Meynard, C. N., V. Devictor, D. Mouillot, W. Thuiller, F. Jiguet and N. Mouquet (2011). "Beyond taxonomic diversity patterns: how do alpha, beta and gamma components of bird functional and phylogenetic diversity respond to environmental gradients across France?" Global Ecology and Biogeography **20**(6): 893-903.

Micheli, F. and B. S. Halpern (2005). "Low functional redundancy in coastal marine assemblages." Ecology Letters **8**(4): 391-400.

Mori, A. S., T. Furukawa and T. Sasaki (2013). "Response diversity determines the resilience of ecosystems to environmental change." Biological Reviews **88**(2): 349-364.

Mouchet, M., F. Guilhaumon, S. Villeger, N. W. H. Mason, J. A. Tomasini and D. Mouillot (2008). "Towards a consensus for calculating dendrogram-based functional diversity indices." Oikos **117**(5): 794-800.

Mouchet, M. A., S. Villeger, N. W. H. Mason and D. Mouillot (2010). "Functional diversity measures: an overview of their redundancy and their ability to discriminate community assembly rules." Functional Ecology **24**(4): 867-876.

Mouillot, D., N. A. J. Graham, S. Villeger, N. W. H. Mason and D. R. Bellwood (2013). "A functional approach reveals community responses to disturbances." Trends in Ecology & Evolution **28**(3): 167-177.

Naeem, S. (1998). "Species redundancy and ecosystem reliability." Conservation Biology **12**(1): 39-45.

Nei, M., Sudhir Kumar (2000). "Molecular evolution and phylogenetics." Oxford university press.

Nelson, M. C., H. G. Morrison, J. Benjamino, S. L. Grim and J. Graf (2014). "Analysis, Optimization and Verification of Illumina-Generated 16S rRNA Gene Amplicon Surveys." Plos One **9**(4).

Newsham, K. K., D. W. Hopkins, L. C. Carvalhais, P. T. Fretwell, S. P. Rushton, A. G. O'Donnell and P. G. Dennis (2016). "Relationship between soil fungal diversity and temperature in the maritime Antarctic." Nature Climate Change **6**(2): 182-+.

Nguyen, N. H., Z. W. Song, S. T. Bates, S. Branco, L. Tedersoo, J. Menke, J. S. Schilling and P. G. Kennedy (2016). "FUNGuild: An open annotation tool for parsing fungal community datasets by ecological guild." Fungal Ecology **20**: 241-248.

Nguyen, N. H., L. J. Williams, J. B. Vincent, A. Stefanski, J. Cavender-Bares, C. Messier, A. Paquette, D. Gravel, P. B. Reich and P. C. Kennedy (2016). "Ectomycorrhizal fungal diversity and saprotrophic fungal diversity are linked to different tree community attributes in a field-based tree experiment." Molecular Ecology **25**(16): 4032-4046.

Nielsen, U. N., E. Ayres, D. H. Wall and R. D. Bardgett (2011). "Soil biodiversity and carbon cycling: a review and synthesis of studies examining diversity-function relationships." European Journal of Soil Science **62**(1): 105-116.

Nilsson, R. H., M. Ryberg, K. Abarenkov, E. Sjokvist and E. Kristiansson (2009). "The ITS region as a target for characterization of fungal communities using emerging sequencing technologies." Fems Microbiology Letters **296**(1): 97-101.

Pavoine, S. and M. B. Bonsall (2011). "Measuring biodiversity to explain community assembly: a unified approach." Biol Rev Camb Philos Soc **86**(4): 792-812.

Pavoine, S., S. Ollier and D. Pontier (2005). "Measuring, diversity from dissimilarities with Rao's quadratic entropy: Are any dissimilarities suitable?" Theoretical Population Biology **67**(4): 231-239.

Perronne, R., F. Munoz, B. Borgya, X. Reboud and S. Gaba (2017). "How to design trait-based analyses of community assembly mechanisms: Insights and guidelines from a literature review." Perspectives in Plant Ecology Evolution and Systematics **25**: 29-44.

Petchey, O. L. and K. J. Gaston (2002). "Functional diversity (FD), species richness and community composition." Ecology Letters **5**(3): 402-411.

Petchey, O. L. and K. J. Gaston (2006). "Functional diversity: back to basics and looking forward." Ecology Letters **9**(6): 741-758.

Pillar, V. D., C. C. Blanco, S. C. Muller, E. E. Sosinski, F. Joner and L. D. S. Duarte (2013). "Functional redundancy and stability in plant communities." Journal of Vegetation Science **24**(5): 963-974.

Pinto, A. J. and L. Raskin (2012). "PCR biases distort bacterial and archaeal community structure in pyrosequencing datasets." PLoS One **7**(8): e43093.

Preston-Mafham, J., L. Boddy and P. F. Randerson (2002). "Analysis of microbial community functional diversity using sole-carbon-source utilisation profiles - a critique." Fems Microbiology Ecology **42**(1): 1-14.

Rao, C. R. (1982). "Diversity and Dissimilarity Coefficients - a Unified Approach." Theoretical Population Biology **21**(1): 24-43.

Ricotta, C. (2005). "Additive partitioning of Rao's quadratic diversity: a hierarchical approach." Ecological Modelling **183**(4): 365-371.

Ricotta, C. (2005). "A note on functional diversity measures." Basic and Applied Ecology **6**(5): 479-486.

Ricotta, C., F. de Bello, M. Moretti, M. Caccianiga, B. E. L. Cerabolini and S. Pavoine (2016). "Measuring the functional redundancy of biological communities: a quantitative guide." Methods in Ecology and Evolution **7**(11): 1386-1395.

Ricotta, C. and L. Szeidl (2009). "Diversity partitioning of Rao's quadratic entropy." Theoretical Population Biology **76**(4): 299-302.

Robeson, M. S., A. J. King, K. R. Freeman, C. W. Birky, Jr., A. P. Martin and S. K. Schmidt (2011). "Soil rotifer communities are extremely diverse globally but spatially autocorrelated locally." Proc Natl Acad Sci U S A **108**(11): 4406-4410.

Sala, O. E., F. S. Chapin, J. J. Armesto, E. Berlow, J. Bloomfield, R. Dirzo, E. Huber-Sanwald, L. F. Huenneke, R. B. Jackson, A. Kinzig, R. Leemans, D. M. Lodge, H. A. Mooney, M. Oesterheld, N. L. Poff, M. T. Sykes, B. H. Walker, M. Walker and D. H. Wall (2000). "Biodiversity - Global biodiversity scenarios for the year 2100." Science **287**(5459): 1770-1774.

Schirmer, M., U. Z. Ijaz, R. D'Amore, N. Hall, W. T. Sloan and C. Quince (2015). "Insight into biases and sequencing errors for amplicon sequencing with the Illumina MiSeq platform." Nucleic Acids Research **43**(6).

Schloss, P. D., S. L. Westcott, T. Ryabin, J. R. Hall, M. Hartmann, E. B. Hollister, R. A. Lesniewski, B. B. Oakley, D. H. Parks, C. J. Robinson, J. W. Sahl, B. Stres, G. G. Thallinger, D. J. Van Horn and C. F. Weber (2009). "Introducing mothur: open-source, platform-independent, community-supported software for describing and comparing microbial communities." Appl Environ Microbiol **75**(23): 7537-7541.

Schmera, D., T. Eros and J. Podani (2009). "A measure for assessing functional diversity in ecological communities." Aquatic Ecology **43**(1): 157-167.

Schmera, D., J. Heino, J. Podani, T. Eros and S. Doledec (2017). "Functional diversity: a review of methodology and current knowledge in freshwater macroinvertebrate research." Hydrobiologia **787**(1): 27-44.

Scrucca, L., M. Fop, T. B. Murphy and A. E. Raftery (2016). "mclust 5: Clustering, Classification and Density Estimation Using Gaussian Finite Mixture Models." R J **8**(1): 289-317.

Sharpe, R. A., N. Bearman, C. R. Thornton, K. Husk and N. J. Osborne (2015). "Indoor fungal diversity and asthma: A meta-analysis and systematic review of risk factors." Journal of Allergy and Clinical Immunology **135**(1): 110-122.

Shelobolina, E. S., M. V. Coppi, A. A. Korenevsky, L. N. DiDonato, S. A. Sullivan, H. Konishi, H. F. Xu, C. Leang, J. E. Butler, B. C. Kim and D. R. Lovley (2007). "Importance of c-type cytochromes for U(VI) reduction by Geobacter sulfurreducens." Bmc Microbiology **7**.

Shimatani, K. (2001). "On the measurement of species diversity incorporating species differences." Oikos **93**(1): 135-147.

Sikkema-Raddatz, B., L. F. Johansson, E. N. de Boer, R. Almomani, L. G. Boven, M. P. van den Berg, K. Y. van Spaendonck-Zwarts, J. P. van Tintelen, R. H. Sijmons, J. D. H. Jongbloed and R. J. Sinke (2013). "Targeted Next-Generation Sequencing can Replace Sanger Sequencing in Clinical Diagnostics." Human Mutation **34**(7): 1035-1042.

Silva, M. C. P. E., B. Schloter-Hai, M. Schloter, J. D. van Elsas and J. F. Salles (2013). "Temporal Dynamics of Abundance and Composition of Nitrogen-Fixing Communities across Agricultural Soils." Plos One **8**(9).

Singh, J. S. (2015). "Microbes: The chief ecological engineers in reinstating equilibrium in degraded ecosystems." Agriculture Ecosystems & Environment **203**: 80-82.

Sogin, M. L., H. G. Morrison, J. A. Huber, D. Mark Welch, S. M. Huse, P. R. Neal, J. M. Arrieta and G. J. Herndl (2006). "Microbial diversity in the deep sea and the underexplored "rare biosphere"." Proceedings of the National Academy of Sciences of the United States of America **103**(32): 12115-12120.

Sokal, R. R. (1958). "A statistical method for evaluating systematic relationship." University of Kansas science bulletin **28**: 1409-1438.

Tajima, F. and M. Nei (1984). "Estimation of Evolutionary Distance between Nucleotide-Sequences." Molecular Biology and Evolution **1**(3): 269-285.

Talbot, J. M., T. D. Bruns, J. W. Taylor, D. P. Smith, S. Branco, S. I. Glassman, S. Erlandson, R. Vilgalys, H. L. Liao, M. E. Smith and K. G. Peay (2014). "Endemism and functional convergence across the North American soil mycobiome." Proc Natl Acad Sci U S A **111**(17): 6341-6346.

Tamura, K. (1992). "Estimation of the Number of Nucleotide Substitutions When There Are Strong Transition-Transversion and G+C-Content Biases." Molecular Biology and Evolution **9**(4): 678-687.

Tamura, K. and M. Nei (1993). "Estimation of the Number of Nucleotide Substitutions in the Control Region of Mitochondrial-DNA in Humans and Chimpanzees." Molecular Biology and Evolution **10**(3): 512-526.

Tedersoo, L., M. Bahram, S. Polme, U. Koljalg, N. S. Yorou, R. Wijesundera, L. V. Ruiz, A. M. Vasco-Palacios, P. Q. Thu, A. Suija, M. E. Smith, C. Sharp, E. Saluveer, A. Saitta, M. Rosas, T. Riit, D. Ratkowsky, K. Pritsch, K. Poldmaa, M. Piepenbring, C. Phosri, M. Peterson, K. Parts, K. Partel, E. Otsing, E. Nouhra, A. L. Njouonkou, R. H. Nilsson, L. N. Morgado, J. Mayor, T. W. May, L. Majuakim, D. J. Lodge, S. S. Lee, K. H. Larsson, P. Kohout, K. Hosaka, I. Hiiesalu, T. W. Henkel, H. Harend, L. D. Guo, A. Greslebin, G. Grelet, J. Geml, G. Gates, W. Dunstan, C. Dunk, R. Drenkhan, J. Dearnaley, A. De Kesel, T. Dang, X. Chen, F. Buegger, F. Q. Brearley, G. Bonito, S. Anslan, S. Abell and K. Abarenkov (2014). "Global diversity and geography of soil fungi." Science **346**(6213): 1078-+.

Tilman, D., P. B. Reich and J. M. H. Knops (2006). "Biodiversity and ecosystem stability in a decade-long grassland experiment." Nature **441**(7093): 629-632.

Tiquia, S. M., L. Y. Wu, S. C. Chong, S. Passovets, D. Xu, Y. Xu and J. Z. Zhou (2004). "Evaluation of 50-mer oligonucleotide arrays for detecting microbial populations in environmental samples." Biotechniques **36**(4): 664-+.

Tromas, N., N. Fortin, L. Bedrani, Y. Terrat, P. Cardoso, D. Bird, C. W. Greer and B. J. Shapiro (2017). "Characterising and predicting cyanobacterial blooms in an 8-year amplicon sequencing time course." Isme Journal **11**(8): 1746-1763.

Tu, Q., H. Yu, Z. He, Y. Deng, L. Wu, J. D. Van Nostrand, A. Zhou, J. Voordeckers, Y. J. Lee, Y. Qin, C. L. Hemme, Z. Shi, K. Xue, T. Yuan, A. Wang and J. Zhou (2014). "GeoChip 4: a functional gene-array-based high-throughput environmental technology for microbial community analysis." Mol Ecol Resour **14**(5): 914-928.

Tu, Q. C., H. Yu, Z. L. He, Y. Deng, L. Y. Wu, J. D. Van Nostrand, A. F. Zhou, J. Voordeckers, Y. J. Lee, Y. J. Qin, C. L. Hemme, Z. Shi, K. Xue, T. Yuan, A. J. Wang and J. Z. Zhou (2014). "GeoChip 4: a functional gene-array-based high-throughput environmental technology for microbial community analysis." Molecular Ecology Resources **14**(5): 914-928.

Tuomisto, H. (2010). "A diversity of beta diversities: straightening up a concept gone awry. Part 1. Defining beta diversity as a function of alpha and gamma diversity." Ecography **33**(1): 2-22.

Van de Peer, Y. (2009). "Phylogenetic inference based on distance methods." The phylogenetic handbook: 142-160.

Villeger, S., N. W. H. Mason and D. Mouillot (2008). "New multidimensional functional diversity indices for a multifaceted framework in functional ecology." Ecology **89**(8): 2290-2301.

Villeger, S. and D. Mouillot (2008). "Additive partitioning of diversity including species differences: a comment on Hardy & Senterre (2007)." Journal of Ecology **96**(5): 845-848.

von Wintzingerode, F., U. B. Gobel and E. Stackebrandt (1997). "Determination of microbial diversity in environmental samples: pitfalls of PCR-based rRNA analysis." Fems Microbiology Reviews **21**(3): 213-229.

Vorosmarty, C. J., P. B. McIntyre, M. O. Gessner, D. Dudgeon, A. Prusevich, P. Green, S. Glidden, S. E. Bunn, C. A. Sullivan, C. R. Liermann and P. M. Davies (2010). "Global threats to human water security and river biodiversity." Nature **467**(7315): 555-561.

Wagg, C., S. F. Bender, F. Widmer and M. G. A. van der Heijden (2014). "Soil biodiversity and soil community composition determine ecosystem multifunctionality." Proceedings of the National Academy of Sciences of the United States of America **111**(14): 5266-5270.

Walker, B., A. Kinzig and J. Langridge (1999). "Plant attribute diversity, resilience, and ecosystem function: The nature and significance of dominant and minor species." Ecosystems **2**(2): 95-113.

Wang, J. T., Y. M. Zheng, H. W. Hu, L. M. Zhang, J. Li and J. Z. He (2015). "Soil pH determines the alpha diversity but not beta diversity of soil fungal community along altitude in a typical Tibetan forest ecosystem." Journal of Soils and Sediments **15**(5): 1224-1232.

Wang, Q., G. M. Garrity, J. M. Tiedje and J. R. Cole (2007). "Naive Bayesian classifier for rapid assignment of rRNA sequences into the new bacterial taxonomy." Applied and Environmental Microbiology **73**(16): 5261-5267.

Wang, Q., G. M. Garrity, J. M. Tiedje and J. R. Cole (2007). "Naive Bayesian classifier for rapid assignment of rRNA sequences into the new bacterial taxonomy." Appl Environ Microbiol **73**(16): 5261-5267.

Wang, Q., J. F. Quensen, 3rd, J. A. Fish, T. K. Lee, Y. Sun, J. M. Tiedje and J. R. Cole (2013). "Ecological patterns of nifH genes in four terrestrial climatic zones explored

with targeted metagenomics using FrameBot, a new informatics tool." MBio **4**(5): e00592-00513.

Whitman, W. B., D. C. Coleman and W. J. Wiebe (1998). "Prokaryotes: the unseen majority." Proc Natl Acad Sci U S A **95**(12): 6578-6583.

Whittaker, R. H. (1960). "Vegetation of the Siskiyou Mountains, Oregon and California." Ecological Monographs **30**(3): 280-338.

Woo, C., C. An, S. Xu, S. M. Yi and N. Yamamoto (2018). "Taxonomic diversity of fungi deposited from the atmosphere." Isme Journal **12**(8): 2051-2060.

Wu, L., C. Wen, Y. Qin, H. Yin, Q. Tu, J. D. Van Nostrand, T. Yuan, M. Yuan, Y. Deng and J. Zhou (2015). "Phasing amplicon sequencing on Illumina Miseq for robust environmental microbial community analysis." BMC Microbiol **15**: 125.

Xie, J. P., L. Y. Wu, J. D. van Nostrand, Z. L. He, Z. M. Lu, H. Yu, J. B. Xiong, X. X. Liu and J. Z. Zhou (2012). "Improvements on environmental DNA extraction and purification procedures for matagenomic analysis." Journal of Central South University **19**(11): 3055-3063.

Xue, K., M. M. Yuan, Z. J. Shi, Y. J. Qin, Y. Deng, L. Cheng, L. Y. Wu, Z. L. He, J. D. Van Nostrand, R. Bracho, S. Natali, E. A. G. Schuur, C. W. Luo, K. T. Konstantinidis, Q. Wang, J. R. Cole, J. M. Tiedje, Y. Q. Luo and J. Z. Zhou (2016). "Tundra soil carbon is vulnerable to rapid microbial decomposition under climate warming." Nature Climate Change **6**(6): 595-+.

Zak, J. C., M. R. Willig, D. L. Moorhead and H. G. Wildman (1994). "Functional Diversity of Microbial Communities - a Quantitative Approach." Soil Biology & Biochemistry **26**(9): 1101-1108.

Zhang, P., W. M. Wu, J. D. Van Nostrand, Y. Deng, Z. L. He, T. Gihring, G. X. Zhang, C. W. Schadt, D. Watson, P. Jardine, C. S. Criddle, S. Brooks, T. L. Marsh, J. M. Tiedje, A. P. Arkin and J. Z. Zhou (2015). "Dynamic Succession of Groundwater Functional Microbial Communities in Response to Emulsified Vegetable Oil Amendment during Sustained In Situ U(VI) Reduction." Applied and Environmental Microbiology **81**(12): 4164-4172.

Zhou, J., M. A. Bruns and J. M. Tiedje (1996). "DNA recovery from soils of diverse composition." Appl Environ Microbiol **62**(2): 316-322.

Zhou, J., Z. He, Y. Yang, Y. Deng, S. G. Tringe and L. Alvarez-Cohen (2015). "High-throughput metagenomic technologies for complex microbial community analysis: open and closed formats." MBio **6**(1).

Zhou, J. Z., Y. Deng, L. N. Shen, C. Q. Wen, Q. Y. Yan, D. L. Ning, Y. J. Qin, K. Xue, L. Y. Wu, Z. L. He, J. W. Voordeckers, J. D. Van Nostrand, V. Buzzard, S. T. Michaletz, B. J. Enquist, M. D. Weiser, M. Kaspari, R. Waide, Y. F. Yang and J. H.

Brown (2016). "Temperature mediates continental-scale diversity of microbes in forest soils." Nature Communications **7**.