UNIVERSITY OF OKLAHOMA

GRADUATE COLLEGE

DEVELOPMENT OF HIGH-THROUGHPUT EXPERIMENTAL AND COMPUTATIONAL TECHNOLOGIES FOR ANALYZING MICROBIAL FUNCTIONS AND INTERACTIONS IN ENVIRONMENTAL METAGENOMES

A DISSERTATION

SUBMITTED TO THE GRADUATE FACULTY

in partial fulfillment of the requirements for the

Degree of

DOCTOR OF PHILOSOPHY

By

ZHOU SHI
Norman, Oklahoma
2017

DEVELOPMENT OF HIGH-THROUGHPUT EXPERIMENTAL AND COMPUTATIONAL TECHNOLOGIES FOR ANALYZING MICROBIAL FUNCTIONS AND INTERACTIONS IN ENVIRONMENTAL METAGENOMES

A DISSERTATION APPROVED FOR THE
DEPARTMENT OF MICROBIOLOGY AND PLANT BIOLOGY

BY

_____
Dr. Lee Krumholz, Chair

_____
Dr. Meijun Zhu

_____
Dr. Sridhar Radhakrishnan

_____
Dr. Yiqi Luo

_____
Dr. Jizhong Zhou

Dedication

I dedicate this dissertation to my wife, Mengting Yuan, whom I met and married during the time period of this work and whose love, faith and support is the greatest gift that I can ever imagine in my life; to my son, Andrew Shi, who is always curious, energetic and wonderful, and is my lasting source of enlightenment and true happiness; to my parents, Yiping Shi and Aiping Zeng, whose selfless love and care transcended the distance and time, and carried me whenever I need; to my aunt, Cindy Shi, whose unwavering support and encouragement instilled in me the spirit of striving for excellence; to my friend, Su Xu, who boarded on the same plane with me to the U.S. seven years ago and then has witnessed all parts of this work as my best friend, and Dian Zou, who passed away in an accident during the time period of this work and is a friend I will forever miss.

Acknowledgement

Pursuing the degree in this program is such a unique and transformative journey to me. As the end of my journey draws near, I would like to take this special opportunity to thank all the people who lifted me up through the past seven years. Without their help and support, I would have been still struggling in the middle of nowhere.

First and foremost, I would like to express my immeasurable gratitude to my advisor, Dr. Jizhong Zhou, for providing me the precious opportunity to start my journey in the first place, and thereafter tremendous support and guidance throughout my entire study and research. The diverse training and research experience I received during this work under his supervision will benefit my future career undoubtedly and persistently.

Also, I am deeply grateful that my advisory committee has been always greatly supportive and beneficial. I wished to thank Dr. Lee Krumholz, my advisory committee chair, for his help in improving my knowledge in a broad range of topics in molecular biology and microbial ecology, and for his guidance and suggestion in scientific writing and academic career development; Dr. Meijun Zhu, for his expert knowledge in validating and improving the mathematical parts of this work and his advice on my professional career from computational perspective; Dr. Yiqi Luo, for his lectures and insights that inspired me to see microbes as a part of the earth system and taught me the essentials in predictive research based on the paradigm of modeling; and Dr. Sridhar Radhakrishnan, for his expertise in computer science and data science, and his humorous and inspiring encouragement on this work and my professional career.

I also wish to express my sincerest appreciation to many of my colleagues at IEG and project collaborators, who provided their kind help and cooperation in every single

project in which I was involved. Their patience in explaining basic knowledge in biology and willingness to answer my absurd questions helped me speed up my adaptation to interdisciplinary research in this program and gain the confidence and experience in the accomplishing this work, especially when I just started it.

Table of contents

List of Tables

List of Figures

Figure 5.6 An example of the overlap between the critical transition and the scale-freeness plateau in a model system (BCI) based on the CLR_PCC method. (a) Changes of scale-freeness with increasing cutoff values. The critical transition region is in yellow which is in between two red dashed lines. The critical threshold is indicated by the blue dashed line. (b) Mean and variance of scale-freeness in a window with the same spanning as the critical transition, whose left side slides from beginning cutoff (0.2) to the beginning of the critical transition (0.672). The blue dashed line indicates those in the critical transition. (c) Permutation (n=9999) test to verify mean of scale-freeness in the critical transition is significantly higher than what from other regions. Means of scale-freeness from permutations are distributed in light blue shape, the original mean of scale-freeness is indicated by the blue dashed line. (b) Permutation (n=9999) test to verify variance of scale-freeness in the critical transition is significantly lower than what from other regions. Variances of scale-freeness from permutations are distributed in light red shape, the original variance of scale-freeness is indicated by the red dashed line. (e) Changes of β value of the model system at the critical threshold in response to the fractions of rewired edges with two different rewiring procedures. Both the procedures randomly change the organization of and preserves the number of edges. 141

Abstract

Microorganisms are ubiquitous on earth, and they interact each other to form communities, which play unique and integral roles in various biochemical processes and functions that are of critical importance in global biogeochemical cycling, human health, energy, climate change, environmental remediation, engineering, industry, and agriculture. However, identification, characterization, and quantification of microbial communities are still limited, due to the extreme diversity and yet-uncultivable nature of a vast majority of microorganisms, and our understanding of microbial communities is further hindered by complex organization and dynamics of interactions among microorganisms. In this work, we developed high-throughput functional gene arrays (FGAs), bioinformatics tools and computational methods for analysis of microbial metagenomes and interactomes to address some of the limitations, whose powerfulness were demonstrated in application studies.

In the beginning of this work, we developed a high-throughput FGA for characterizing a specific group of microorganisms - plant growth promoting microorganisms (PGPMs). PGPMs can promote plant growth and suppress disease directly and/or indirectly by enhancing soil fertility and plant resistance to biotic and abiotic stresses, thus may contribute to the success of invasive plants over native species. However, PGPMs are highly diverse in terms of both species richness and plant promoting mechanisms. Therefore, it is difficult to study the PGPMs changes along with environment shifts, and their subsequent impacts on plant performance and ecosystem functioning. The developed high-throughput FGA, termed Plant Associated Beneficial Microorganism Chip (PABMC), focused on functional genes from PGPMs that are beneficial to plants.

A total of 3,870 probes covering 34 functional gene families were designed in PABMC, including six categories: plant growth-promoting hormones, plant pathogen resistance, antibiotics, antioxidants, drought tolerance, and secondary benefits (e.g. elicitor of plant immune defense response). Computational analysis showed that ~98% of the probes were highly specific at the species or strain level. The PABMC was also applied to investigate PGPMs' responses to *Ageratina adenophora* (*A. adenophora*) invasion in a natural grassland, and showed *A. adenophora* invasion increased the alpha diversity and shifted the composition of PGPM communities compared with what from the native site. The PABMC uncovered changes in abundance of a key gene related to drought tolerance, pathogen resistance, antibiotic biosynthesis, and antioxidant biosynthesis, due to *A. adenophora* invasion. These changes may promote the survival and growth of *A. adenophora* over native species in the site we studied.

Next, we developed GeoChip 5.0, and advanced the FGA based metagenomics technology to a new level of comprehensiveness, for analyzing complex microbial communities. GeoChip 5.0 was based on Agilent platform, with two formats. The smaller format contained 60K probes (GeoChip 5.0S), majorly covering probes from carbon (C), nitrogen (N), sulphur (S), and phosphorus (P) cyclings and energy metabolism probes. The larger format (GeoChip 5.0M) contained all probes in GeoChip 5.0S and expanded to antibiotic resistance, metal resistance/reduction, organic contaminant remediation, stress responses, pathogenesis, soil beneficial microbes, soil pathogens, and virulence. GeoChip 5.0M contains 161,961 probes covering approximately 370,000 representative coding sequences from 1,447 functional gene families. These genes were derived from functionally divergent broad taxonomic

groups, including bacteria (2,721 genera), archaea (101 genera), fungi (297 genera), protists (219), and viruses (167 genera, mainly phages). Both computational and experimental evaluation indicated that all designed probes were highly specific to their corresponding targets. Sensitivity tests revealed that as little as 0.05 ng of pure culture DNAs was detectable within 1 μg of complex soil community DNA as background, suggesting that the Agilent platform-based GeoChip is extremely sensitive. Additionally, very strong quantitative linear relationships were obtained between signal intensity and pure genomic DNAs or soil DNAs. Application of the designed FGAs to a contaminated groundwater with very low biomass indicated that environmental contaminants (majorly, heavy metals) had significant impacts on the biodiversity of microbial communities.

Since next generation sequencing (NGS) technology has revolutionized metagenomics and microbial ecology studies, immense improvements made in sequencing speed, throughput, and cost. However, NGS technology also produces a formidable number of raw reads which poses computational challenges, especially for analyzing deep shotgun metagenomics sequencing data. To tackle some of the challenges, we present an Ecological Function oriented Metagenomic Analysis Pipeline (EcoFun-MAP), to facilitate analysis of shotgun metagenomic sequencing data in microbial ecology studies. The EcoFun-MAP consists of reference databases of different data structures, with a selective coverage of functional genes that are important to ecological functions. Meanwhile, multiple predefined data analysis workflows were built on the databases with most updated bioinformatics tools. Furthermore, the EcoFun-MAP was implemented and deployed on High-Performance Computing (HPC) infrastructure with

high accessible and easy-to-use interfaces. In our evaluation, the EcoFun-MAP was found to be fast (multi-million reads/min.) and highly scalable, and capable of addressing disparate needs for accuracy and precision. In addition, we showcase the effectiveness of the EcoFun-MAP by applying it to reveal differences among metagenomes from underground water samples, and provide insights to link the metagenomic differences with distinctive levels of contaminants.

To extend an emerging dimension of microbial community analysis, that is the analysis of complex microbial interactions, we provided a generalized Brody distribution (GBD) based Random Matrix Theory approach (GBD-RMT approach) for inferring microbial data association networks. The GBD-RMT approach addresses several limitations of a previous Random Matrix Theory (RMT)-based approach in the capability of detection and interpretability of detected thresholds. The GBD-RMT approach is capable of quantitatively characterizing the dynamics of Nearest Neighboring Spacing Distribution (NNSD) of eigenvalues against candidate thresholds, and detecting both the critical transitions and thresholds in NNSD dynamics using trend analysis. In our evaluation, the GBD-RMT approach successfully detected the critical thresholds in all of the numerically simulated and real datasets, including those for which the previous method failed. It also had higher detection resolution, and gained higher confidence and interpretability in detected critical thresholds. Meanwhile, the GBD-RMT approach integrated improvements for detecting more types of data association and reducing compositional data bias. In addition, the GBD-RMT approach uncovered a remarkable overlap between the critical transitions and the plateaus of scale-freeness from the

inferred networks, and the overlap is showed to be statistically significant and universal in complex biological systems in our analysis.

All the developed technologies and computational methods in this work provided powerful and up-to-date means for analyzing complex metagenomes, and should be ready to serve for improving our understanding of microbial communities in the studies of microbial ecology and global change biology.

Keywords: High-throughput metagenomics technology; Functional Gene Array; GeoChip; shotgun metagenome sequencing; bioinformatics tools; computational biology; environmental metagenomics; microbial functions; microbial interactions; plant invasion; underground water contamination; microbial ecological network; Random Matrix Theory; data association networks.

Chapter 1: Introduction

1.1 Limitations of culture-dependent methods in microbial ecology research

Microorganisms are almost ubiquitous in the biosphere, and their existence and functions, to a large extent, shape the biogeochemical cycling of essential elements for life on earth. The science about microbial diversity, community composition, function, interaction, succession, and responses to stimuli in various ecosystems greatly benefit our survival through promoting our exploration of nature, and the development of agriculture, medical care, waste treatment, etc. (Curtis, Head et al. 2003, Zhou, Deng et al. 2014). However, the detection, identification, and characterization of microorganisms in the environment has been challenged by their tiny body size, enormous diversity, versatile and variable functional roles, and complicated interactions amongst themselves and with their biotic and abiotic surroundings (Gans, Wolinsky et al. 2005, Schloss and Handelsman 2006, Sogin, Morrison et al. 2006, Roesch, Fulthorpe et al. 2007, Zhou, He et al. 2015).

Before the popularization of high-throughput metagenomic technologies, microbial ecology research solely depended on cultivation-dependent microbiology. The identification and characterization of microbial taxa were based on morphology (of cells or colonies) and physiology tests after separation of the organism of interest from a vast background community, followed by cultivation. Limitations in such methodology include several. First, 99% of the microorganism discovered are yet uncultivable in known media (Rappe and Giovannoni 2003), posing the question of losing a majority of the diversity and potential functional roles in observations of microbial communities in niche-rich environments such as soil, ocean, and even human body (Whitman, Coleman

et al. 1998, Kallmeyer, Pockalny et al. 2012). Second, although the morphology description and physiological tests are necessary to characterize a specific organism, these methods are low in efficiency to screen communities of microorganisms in defined habitats. Third, the added difficulty in co-culturing different microorganisms hinders the possibility to study microbial interactions (Fuhrman 2009, Zhou, Deng et al. 2010). Fourth, the laboratory culture media can hardly simulate natural environments in all aspects, biasing the estimation of *in situ* conditions of microorganisms (Fitter, Gilligan et al. 2005, Levin 2006). In a word, using only culture-dependent methods, microbial ecology studies can hardly be comprehensive and conclusive.

1.2 Overview of high-throughput metagenomic technologies

Since the last decade in the 20th century, cultivation-independent detection of environmental microorganisms has been developed and popular in microbial community profiling. These methods took advantage of the discoveries of, and the molecular techniques to track and distinguish, multiple biomarkers. For example, PCR amplification-enabled sequencing of 16S rRNA genes (Schmidt, DeLong et al. 1991), amplified ribosomal DNA restriction analysis (Massol-Deya, Weller et al. 1997), denaturing gradient gel electrophoresis (Muyzer, De Waal et al. 1993) and terminal restriction fragment length polymorphism (Liu, Marsh et al. 1997) can use the ribosomal RNA as well as functional gene sequences as biomarkers, phospholipid fatty acid analysis (Frostegård, Tunlid et al. 2011) uses the molecular structure of phospholipid fatty acids in cell membrane as biomarkers, while Biolog EcoPlates (Hadwin, Del Rio et al. 2006) utilize the profile of carbon and nitrogen metabolism potentials as biomarkers, to survey the taxonomic or functional group composition of

microorganisms present in the environmental samples. During the last decade, more efficient, extremely high-throughput technologies that detect several thousand biomarkers, or even the whole genome/transcriptome/metabolome, were developed and became cost-effective, which revolutionized microbial ecology research by enabling deep surveys of the microbial "dark matters", as well as high-resolution comparisons of different communities. These, methods, including next-generation sequencing of DNA/RNA (Venter, Remington et al. 2004, Caporaso, Lauber et al. 2012, Loman, Misra et al. 2012, Weinstock 2012), PhyloChip (Hazen, Dubinsky et al. 2010), GeoChip (He, Deng et al. 2010, Tu, Yu et al. 2014), mass spectrometry-based proteomics (Ram, VerBerkmoes et al. 2005), and metabolite analysis (Cui, Lewis et al. 2008), have been applied to microbial samples from diverse ecosystems to address a wide range of microbial ecology questions.

Among these technologies, DNA-based high-throughput sequencing and microarray technologies are most broadly used as tools to answer the questions of "who is there" and "what they are capable of". Both tools have their own advantage and drawbacks, which were compared and discussed in details in terms of the two major categories of high-throughput technologies: open and closed formats (Vieites, Guazzaroni et al. 2009, Roh, Abell et al. 2010, Zhou, He et al. 2015). "Open format" technologies refer to those do not require a priori profiling of the target aspects of sample community, such as high-throughput sequencing, fingerprinting, and mass spectrometry-based proteomic and metabolomic approaches. Results from these technologies frequently contain outcomes that are not previously described, such as new sequences, pathways, etc., enabling discovery of novel species, yet often have the problem of undersampling thus

loss of low-abundant species. On the contrary, "closed format" technologies use defined profiling based on previous knowledge to detect the existence and/or abundance of the species, or the level of realized functions, such as microarrays, Biolog EcoPlates, and quantitative PCR. These technologies do not recover novel molecular information from the sample community but are usually more sensitive to rare members and can be more quantitative compared with open format applications.

Thus, open-format metagenomic sequencing and closed-format microarray could be complementary tools, whose combination could comprehensively profile environmental microbial samples in high resolution (Zhou, He et al. 2015). Such profiling will aid to address fundamental microbial ecology questions, such as community diversity and succession, as well as link the molecular information to ecosystem functions. This also enables the exploration of complex microbial interactions through network inference based on co-occurrence and abundance patterns, from which keystone species with ecological importance could be identified.

1.3 High-throughput sequencing and challenges

Next generation sequencing (NGS), or high-throughput sequencing technology, utilizes the sequencing by synthesis (SBB) approach to track the identity of the fluorescently labeled nucleotide during its addition to the nucleotide chain. It allows massively paralleled detection of millions of sequences in a single run, revolutionized the sequencing ability of classic Sanger chain-termination method. NGS technology includes several platforms, such as Illumina, Roche 454, SOLiD sequencing, among which Illumina has become the most high-throughput, cost- and time-effective, and popular one in many research areas including microbial ecology (Metzker 2010,

Shokralla, Spall et al. 2012). Equipped with the ability to get in-depth profiling of microbial community from the environmental sample, and discover new microbial species and functional modules, NGS technology has greatly facilitated microbial ecology studies by unraveling previously hidden information in the microbial community "black box"(Tiedje, Asuming-Brempong et al. 1999).

NGS of DNA has two major applications: amplicon sequencing and metagenomics shotgun sequencing (Scholz, Lo et al. 2012). Amplicon sequencing surveys the PCR amplicons-based library of phylogenetic marker genes (e.g., 16S and 18S rRNA, ITS) or functional genes (e.g., *nifH, amoA*), but often introduces biased estimation during the required PCR amplification. Metagenomic shotgun sequencing avoids PCR and can recover sequence fragments from, theoretically, all over the genomes of the DNA samples. However, very deep sequencing, which is necessary to reveal microbial community functional composition in samples from complex systems, such as soil, post great computational challenge in terms of both data storage, transfer, management, and the retrieval of biologically meaningful information from the sequences (Scholz, Lo et al. 2012). Although the rapid development of bioinformatic tools and databases partly enabled the decipher of complex genetic codes, the under-standardized database, varied algorithms for each analysis step, and the need for intensive coding still create barriers for microbial ecologists to efficiently and accurately clean, analyze, and take information from the metagenomic sequencing output. Specifically, developing tools that provide efficient and accessible solutions for functional analysis of shotgun metagenomics data, with the focus on linking functional composition of the microbial

community to ecological functions and geochemical processes, is urgently demanding (Gonzalez and Knight 2012, Scholz, Lo et al. 2012).

1.4 High-throughput DNA microarray and challenges

As an alternative to NGS, DNA microarray technology has also been advanced and broadly used in microbial ecology studies to quantify the relative amount of large numbers of sequences of interest simultaneously (Zhou, He et al. 2015). The capacity (the number of different spots) of microarray slide kept increasing during the past two decades, allowing the detection of thus many different sequences in one hybridization. Since microarray is pre-designed per existing database, its output contains only the signal intensities (abundance) of known oligos, keeping the size of data relatively low. The processing of signal intensity values is usually standardized, even done automatically in developers' software, hence time-saving. Most importantly, microarrays are highly sensitive in capturing low-abundance and in focal species, while efficiently discard any irrelevant information. Different types of microarrays are customized based on scientific questions. For example, PhyloChip uses taxonomic marker sequences to depict the composition of the microbial community, and GeoChip discloses the metric of biogeochemically important microbial functional genes, both widely applied on microbial communities from various habitats (Zhou, He et al. 2015). With the increasing size of sequence database, one challenge of microarray technology is to optimize the probe design to take advantage of the large capacity of the array for more accurate detection of more genes from more complex community. First, the coverage of sequences should be high to capture as many as possible fragments in the community that matches certain function or taxa. Second, the probe sequences should

be well distinguishable from each other to avoid non-specific hybridization. Third, multiple control probes should be designed and able to be used as the reference in data normalization. The second challenge for microarray technology is to develop easily accessible computer software that allows the quick and accurate design of customized array for different suits of scientific questions and hypothesis testing. All these tasks require heavy computational resources regarding the size and update rate of the current database.

1.5 Inference of microbial association networks

In almost any environment, microorganism cells are interacting with each other, or interconnected through various biotic and abiotic environmental factors (Atlas and Bartha 1986, Whipps 2001). For example, inside the sewage treatment active sludge, different types of microorganism form mutualism through substrate chain. Such interactions greatly shape the ecosystem functioning of the microbial community and represent another important dimension, besides the diversity and abundance, of a microbial community (Zhou, Deng et al. 2010, Widder, Besemer et al. 2014, Shi, Nuccio et al. 2016). In complex systems, such as soil, the interactions of microbes are extremely hard to observe, characterize and validate. With the assistance of high-throughput metagenomic technologies and the accumulation of microbial community richness, abundance, and functional profiling records, it becomes increasingly attractive to find ways to detect and quantify the community interactions directly from these data without prior knowledge of the relationships.

The data association network inference is among a number of methods that have been developed for characterizing microbial community interactions based on abundance

observations (Faust and Raes 2012). It offers the opportunity to examine meaningful interactions in large communities, although is not able to determine the causality between given pairs of variables. Two important processes in the inference of data association networks are 1) detection of data association; and 2) selection of threshold of data association strength (Faust and Raes 2012). Different methods can be used in these two steps, and the corresponding outcomes largely influence the properties of constructed network. Therefore, these methods should be carefully evaluated and validated before using them in drawing any biological conclusion.

The detection of data association requires a fast, general and comprehensive technique for detecting and quantifying random variable associating patterns and strength. While earlier studies tended to only explore the linear relationships indicated by the Pearson Correlation Coefficient(Pearson 1901) (PCC), emerging studies applied various data association techniques to recover more complex data association (e.g. nonlinear dependence), which is broadly existed among ecological interactions (e.g., predation, competition and mutualism), extending the scope and precision of interactions that data association network can recover.

Until recently, most studies have selected the association strength threshold empirically, so the constructed networks are inevitably subject to artificial deviations. To solve the problem, a random matrix theory (RMT)-based approach is developed to automatically and objectively detect such a threshold. RMT has been a powerful tool for identifying and modeling the phase transitions and dynamics with disorder and noise in complex systems, including biological systems. The applicability of the RMT in biological systems has been previously demonstrated for inferring metabolic, protein, functional

gene and microbial ecological networks (Luo, Zhong et al. 2006, Luo, Zhong et al. 2006, Luo, Yang et al. 2007, Zhou, Deng et al. 2010, Zhou, Deng et al. 2011), and an RMT-based molecular ecological networks analysis pipeline (MENAP) constructed by Deng et al (Deng, Jiang et al. 2012) has been used to computationally facilitate the network inference. However, the current approach has several limitations. First, the MENAP is limited in detecting data associations other than linear correlation, as it relies on the PCC. Second, the MENAP doesn't have any preprocessing step to remove compositional data bias. Third, the MENAP occasionally failed to detect critical thresholds occasionally. In addition, the MENAP calls critical threshold on each candidate cutoff without telling how good it is for the threshold, which is a lack of quantitative assessment, and made the inferred networks less interpretable. Therefore, for improving the inference of microbial data association networks, those limitations need to be addressed.

1.6 Objectives of this study

This dissertation aimed at addressing some of the technological and computational needs in two metagenomic methods, the functional gene array (FGA) and NGS for microbial community profiling, and developing algorithms for characterizing the microbial interactomes based on the data generated from the FGA and NGS platforms. Following summarized research focus of each chapter.

In Chapter 2, we developed the Plant Associated Beneficial Microorganism Chip (PABMC), which the first high-throughput functional gene array (FGA) focusing on characterizing genes benefiting plants from plant growth promoting microorganisms (PGPMs). In the PABMC, a total of 3,870 probes were designed and computationally

verified to be highly specific, which covered 34 functional gene families from six selected major functional categories, including plant growth-promoting hormones, plant pathogen resistance, antibiotics, antioxidants, drought tolerance, and others (e.g. elicitor of plant immune defense response). Meanwhile, the effectiveness of the PABMC was demonstrated in an application study to investigate PGPMs' responses to *Ageratina adenophora (A. adenophora)* invasion in a natural grassland. The application found the changes in diversity, composition, and abundances of key genes in the microbial communities, which may promote the survival and growth of *A. adenophora* over native species on the site.

In Chapter 3, we further developed a new generation of comprehensive and high-density FGA, GeoChip 5.0, based on Agilent platform, for tackling challenges in the representation, specificity, sensitivity and quantitation in analyses of complex microbial communities. The full version of GeoChip 5.0 contains 161,961 probes covering approximately 370,000 representative coding sequences from 1,447 functional gene families that are involved in a diverse range of physiochemical and biogeochemical processes, for example, carbon and nitrogen cycling, contamination remediation, and antibiotic resistance. These genes were derived from functionally divergent broad taxonomic groups, including bacteria, archaea, fungi, protists, and viruses. Both computational and experimental evaluations were conducted, proving that GeoChip 5.0 is highly specific, sensitive, and quantitative. Application of GeoChip 5.0 to contaminated groundwater samples indicated that it is effective and efficient in detecting the responses of microbial functional gene abundances and diversity to heavy metal contaminants, demonstrating its potential in promoting researches in human

health, agriculture, energy, climate change, ecosystem management, and environmental restoration.

Chapter 4 presents an Ecological Function-oriented Metagenomics Analysis Pipeline (EcoFun-MAP), designed and developed to ease functional analyses of shotgun metagenome sequencing data derived from microbial ecology studies. The EcoFun-MAP consists of reference databases of different types that enable use of bioinformatics tools with distinctive features, and has a selective coverage of functional genes that are important to ecological functions. Meanwhile, multiple predefined data analysis workflows were built on the databases with most updated bioinformatics tools, which allows to processing input sequencing reads, assign them to genes that are important to ecological functions. Furthermore, the EcoFun-MAP was implemented and deployed on High-Performance Computing (HPC) infrastructure with high accessible and easy-to-use interfaces. In our evaluation, the EcoFun-MAP was found to be fast (multi-million reads/min.) and highly scalable, in the meantime accurate and precise. In addition, we showcase the effectiveness of the EcoFun-MAP by applying it to reveal differences among metagenomes from underground water samples and provide insights to link the metagenomic differences with distinctive levels of contaminants.

Chapter 5 a generalized Brody distribution (GBD) based Random Matrix Theory approach (GBD-RMT approach) for inferring microbial data association networks. The GBD-RMT approach acquires the GBD unifying Wigner-Dyson and Poisson distribution with one single parameter, β, which can be used as a quantitative indicator of the transition progress of the NNSD. Maximum Likelihood Estimation (MLE) based method was used for obtaining the best estimation for the β. Meanwhile, the critical

transitions and thresholds were detected using trend analysis on the β dynamics generated from the snapshots of a series of data association matrix reductions with cutoff values from low to high. In the evaluation of the GBD-RMT approach, both in silico and real datasets were used for demonstrating the effectiveness of the approach. Comparisons were also made between the GBD-RMT approach and the previous approach (Luo, Zhong et al. 2006, Luo, Yang et al. 2007). In addition, the GBD-RMT approach was used for uncovering a remarkable linkage between the critical transition of β series and the plateau of scale-freeness from the inferred networks, and the linkage is showed to be statistically significant and universal in complex biological systems in our analysis.

The conclusion chapter summarized the development of the two high-throughput FGAs for the specific and general purpose, respectively, the bioinformatic pipeline for function-oriented analysis of shotgun metagenome sequencing data and the computational approach for inferring data association networks. Overall, the work in this dissertation offered new and up-to-date technological and computational resources for advancing metagenomics studies in microbial ecology.

Chapter 2: Development of a Functional Gene Array to Characterize Plant Growth Promoting Microorganisms Beneficial to Plants

2.1 Abstract

Plant growth promoting microorganisms (PGPMs) can promote plant growth and suppress disease directly and/or indirectly by enhancing soil fertility and plant resistance to biotic and abiotic stresses thus may contribute to the success of invasive plants over native species. However, PGPM is highly diverse in terms of both species richness and plant promoting mechanisms. Therefore, it is difficult to study the PGPMs changes along with environment shifts, and their subsequent impacts on plant performance and ecosystem functioning. Here we developed a microarray focusing on functional genes from PGPMs that are beneficial to plants, termed Plant Associated Beneficial Microorganism Chip (PABMC), to investigate soil PGPMs' responses to the invasive plant species *Ageratina adenophora (A. adenophora)* invasion in a natural grassland. A total of 3,870 probes covering 34 functional gene families were designed in PABMC, including six categories: plant growth-promoting hormones, plant pathogen resistance, antibiotics, antioxidants, drought tolerance, and secondary benefits (e.g. elicitor of plant immune defense response). Computational analysis showed that ~98% of the probes were highly specific at the species or strain level. By applying PABMC to soil, we found that *A. adenophora* invasion increased the alpha diversity and shifted the composition of PGPM communities compared with what from the native site. The abundance of a key gene related to drought tolerance (*tre_arc*) was significantly increased by the invasion, while those for pathogen resistance (*sid_arc* and *sid_fun*) were significantly decreased. Different directions of significant changes were

observed in response to the *A. adenophora* invasion, for both antibiotic biosynthesis, in which abundances increased in two genes (*lgrD* and *pabA*) and decreased in three (*lmbA*, *phzF* and *strR*), and antioxidant biosynthesis, in which abundances increased in one gene (*per_bac*) and decreased in two (*cat_arc* and *sod_nickel*). These changes may promote the survival and growth of *A. adenophora* over native species in the site we studied. In summary, the PABMC provides a novel and high-throughput tool to characterize soil PGPM communities and was proved to be effective when applying in investigating PGPM changes under *A. adenophora* invasion.

2.2 Introduction

Plant root exudates and plant debris incorporated into soils cooperatively affect soil microbial community composition, structure, diversity and functions (Grayston, Wang et al. 1998, Garbeva, Van Veen et al. 2004, el Zahar Haichar, Marol et al. 2008, Berendsen, Pieterse et al. 2012, Turner, Ramakrishnan et al. 2013, Chaparro, Badri et al. 2014, Shi, Nuccio et al. 2016). Meanwhile, soil microorganisms have detrimental, neutral or beneficial influences on plant growth and survival (Stacey and Keen 1995, Barka and Clément 2008, Van Der Heijden, Bardgett et al. 2008, Mendes, Kruijt et al. 2011), though the nature of their influences may change according to plant types and environmental niches. Plant beneficial microorganisms in soil have been well reported to endow various benefits for plant health, contributing to the success of the beneficiary plants in natural or agricultural ecosystems (Davison 1988). For example, soil PGPMs can promote plant growth through releasing plant hormones (e.g. auxins, cytokinins, gibberellins, ethylene and abscisic acid) (Frankenberger Jr and Arshad 1995) or plant hormone precursors (e.g. 1-aminocyclopropane-1-carboxylate) (Lugtenberg and Kamilova 2009), prevent deleterious effects of soil-borne pathogens through generating siderophores (i.e. small iron-binding molecules) (Kloepper, Leong et al. 1980) or antibiotics (Glick 1995) productions, and assist plants in tolerating drought through the regulation of aquaporins to improve soil water status (Maurel 1997). PGPMs can also alter soil properties through solubilizing nutrients, reinforcing resistance of plants to stress, stabilizing soil aggregates, and improving soil structure (Rodríguez and Fraga 1999).

PGPMs were reported to interact with invasive plants differently from their interactions with native plant species (Klironomos 2002, Callaway, Thelen et al. 2004, Van Der Heijden, Bardgett et al. 2008, Rout and Callaway 2009). They may mediate plant invasions directly or indirectly through diverse mechanisms including enemy escape (Klironomos 2002), allelopathic weapon (Cipollini, Rigsby et al. 2012), local pathogen accumulation (Eppinga, Rietkerk et al. 2006, Mangla and Callaway 2008), reinforced mutualism (Reinhart and Callaway 2004), native mutualism interruption (Stinson, Campbell et al. 2006, Callaway, Cipollini et al. 2008) and soil dynamics alteration (Ehrenfeld 2003). For example, compared with native species, some exotic invasive plant species were less suppressed by soil-borne pathogens (Van Grunsven, Van Der Putten et al. 2007), promoted certain soil pathogens to impede seedling growth of the native plant species (Mangla and Callaway 2008), and disrupted native mutualistic plant-microbe interaction (Stinson, Campbell et al. 2006). Therefore, understanding the interaction between PGPMs and plant invasions is essential to prevent further exotic invasions and facilitate the restoration of invaded ecosystems. There has been persistent interest (Callaway and Aschehoug 2000, Klironomos 2002, Callaway, Thelen et al. 2004, Batten, Scow et al. 2006, Broz, Manter et al. 2007, Rout and Callaway 2009, Lorenzo, Pereira et al. 2013, Maron, Klironomos et al. 2014, Carey, Beman et al. 2015, Kowalski, Bacon et al. 2015, Gornish, Fierer et al. 2016, McLeod, Cleveland et al. 2016) in investigating roles of soil microbiota in the spread of invasive species in native ecosystems. However, most of them used traditional methods (e.g. phospholipid fatty acid analysis) with low resolution (citation) or only focused on a few microbial species or mechanisms. Comprehensive coverage of PGPM or insights into responses of PGPM

metagenomes to plant invasions is still lacked and our understanding of interactions between PGPMs and invasive plants was still limited.

It remains challenging to characterize soil microbial communities due to the enormous diversity and as-yet-uncultivated nature of the majority of microorganisms (Whitman, Coleman et al. 1998, Gans, Wolinsky et al. 2005, Schloss and Handelsman 2006, Roesch, Fulthorpe et al. 2007). Characterizing PGPMs is even more difficult because they are highly diverse in plant promoting mechanisms, and usually less abundant in the microbial community. Functional gene array-based technologies, such as GeoChip, have been shown to be as reliable and comprehensive tools to analyze the functional diversity, composition and structure of microbial communities (He, Gentry et al. 2007, Zhou, Kang et al. 2008, Van Nostrand, Wu et al. 2009, Waldron, Wu et al. 2009, Wang, Zhou et al. 2009, He, Deng et al. 2010, He, Xu et al. 2010, Lu, He et al. 2012, Trivedi, He et al. 2012, Zhou, Liu et al. 2013, Tu, Yu et al. 2014). It can harness the unique or conservative regions of the genes encoding key enzymes involved in the synthesis of the metabolites that can be used as indicators to detect and identify gene hosts. Since most PGPMs benefit plant growth and survival through distinctive metabolites, it is possible to use functional gene array to detect microbial functional genes beneficial to plants. To our knowledge, however, no functional gene array has been developed so far to target PGPMs functional genes specifically.

In this study, we developed a specific functional gene array, termed Plant Associated Beneficial Microorganism Chip (PABMC), focused on key functional gene families involved in plant beneficial metabolite synthesis for investigating PGPMs. We computationally evaluated and verified the specificity of the PABMC based on the

sequence identity, continuous matching stretch, and hybridization energy. The developed PABMC was used for study changes of PGPMs along with different intensities of plant invasion in a natural grassland in southeast China where the rapid expansion of *Ageratina adenophora* (*A. adenophora*) has been occurring, and in the most serious cases, transforming the diverse local community into monoculture and posing a serious threat to native biodiversity and productivity. As demonstrated in this study, the PABMC provides an effective high-throughput tool for characterizing microbiomes of PGPM and obtaining insights into their diversity, composition, and structure.

2.3 Materials and methods

2.3.1 Designing and selecting oligonucleotide probes for the PABMC

Oligonucleotide probes (50-mers) targeting microbial genes benefiting plants were designed and selected for the PABMC based on a scheme (**Figure 2.1**) that has been used and validated for efficient functional gene array development (He, Gentry et al. 2007, He, Deng et al. 2010, Tu, Yu et al. 2014). First, functional gene families of interest were selected from the literature, including those that play crucial roles in the synthesis of metabolites benefiting plants through pathogen resistance, plant hormone promotion, antibiotic activity, stress tolerance and other processes. Keywords related to these gene families were identified, and keyword-based queries were manually crafted and submitted for protein sequence retrieval from the NCBI online public databases (i.e. GenBank). Next, seed sequences were manually chosen to build profile hidden Markov models and verify coding sequence (CD) candidates using HMMER 2.3.2 (Eddy 1998). Oligonucleotide probes were then designed to target corresponding nucleotide

sequences of the verified sequences using CommOligo 2.0 (Li, He et al. 2005). Each of the designed probes was further searched against NCBI *nt* and *env_nt* databases using BLAST programs in order to validate the specificity. Best probes from all valid ones were selected and then synthesized onto microarrays by Roche NimbleGen (Madison, WI).

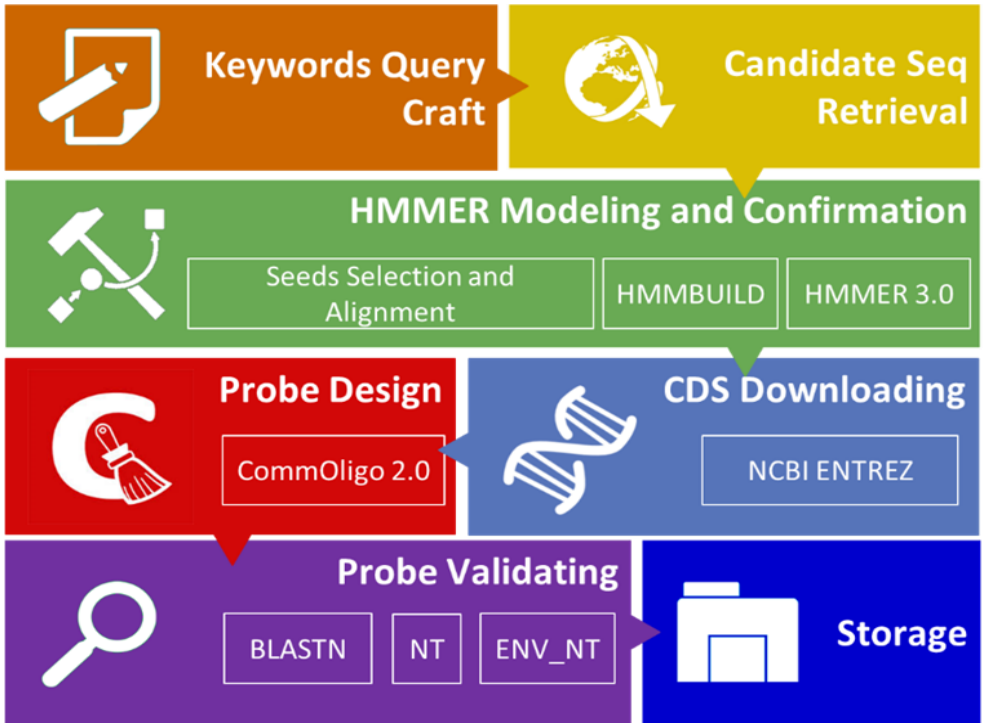


Figure 2.1 The scheme of the automated workflow constructing the PABMC. Similar procedure and criteria as described for GeoChip 3 (He et al., 2010) has been used for sequence retrieval and probe design and selection.

2.3.2 Sample collection, DNA preparation, and microarray hybridization

Nine bulk soil samples were collected from an *A. adenophora* invaded region in Yunnan province, China in December 2010. As shown in the sampling map, **Figure S 1**), three replicated samples (A1 - A3) were collected from center of a patch where *A. adenophora* was dominant (coverage of *A. adenophora* ≥ 60%, age of invasion ≥ 5 years); three replicated samples (AX1 - AX3) were collected from a mixed region

around the edge of the patch where *A. adenophora* and native plants co-existed (coverage of *A. adenophora* is between 10% to 30%, coverage of native plants is between 30% to 50%); and three replicated samples (N1 - N3) were collected from area outside the patch which was not invaded by *A. adenophora* but dominated by native plants (coverage of native plants ≥ 40%). The DNA preparation and microarray hybridization in study used a procedure that was described previously in details (Tu, Yu et al. 2014). Briefly, soil DNA was extracted and purified using previously described methods (Zhou, Bruns et al. 1996). The DNA (1.5 μg) was measured by PicoGreen (Ahn, Costa et al. 1996) and then labeled with Cy-3 and nucleotides (Wu, Liu et al. 2006). After labeling, the DNA was purified and evaluated using a QIA quick purification kit (Qiagen) and NanoDrop (NanoDrop Technologies Inc.), respectively. Next, the DNA was dried and rehydrated with 2.68 μL sample tracking control buffer, and then were incubated, vortexed, and then centrifuged. The samples was mixed with hybridization buffer (7.32 μL) and 2.8% Cy5-labeled CORS target (Tu, Yu et al. 2014). The samples (6.8 μL) were then loaded to the array and hybridized at 42°C approximately 16 h with mixing (Tu, Yu et al. 2014).

2.3.3 Microarray data pre-processing

The hybridized microarray slides were scanned and imaged using a NimbleGen MS 200 Microarray Scanner (Lu, He et al. 2012). Resulting images were then gridded using NimbleScan software (Roche NimbleGen) with a prepared gridding file. In order to remove noise and obtain more reliable microarray data for further analyses, probes with the coefficient of variance (CV) greater than 0.8 were removed. Remaining probes were considered positive if their signal-to-noise ratio (SNR; (probe signal-

background)/background SD;) was at least 2 as previously described (Cui, Lewis et al. 2008). Signal intensity of each spot across all arrays was normalized to the same level with the mean signals of pre-spiked CORS probes (Liang, He et al. 2010). All hybridization data are available at the Institute for Environmental Genomics, University of Oklahoma (http://ieg.ou.edu/4download/).

2.3.4 Statistical analysis

Preprocessed microarray data obtained from each environment sample were used for statistical analyses by the vegan package in R 2.9.1 (Team 2012). Plant beneficial gene diversity was calculated using functional gene richness, Simpson's index, and Shannon-Wiener's diversity index. Non-metric Multidimensional Scaling (NMDS) analysis was used to determine the overall changes in the occurrence and distribution of plant beneficial genes in each microbial community (Zhou, Kang et al. 2008). Three different non-parametric analyses for multivariate data were performed to measure the overall differences of community functional gene structure among samples from regions suffering different invasion levels: 1) analysis of similarities (ANOSIM) (Clarke and Ainsworth 1993), 2) non-parametric multivariate analysis of variance using distance matrices (ADONIS) (Anderson 2001), and 3) multi-response permutation procedure (MRPP) (McCune, Grace et al. 2002, Mielke and Berry 2007). Bray-Curtis similarity index was used to calculate the distance matrix for all three methods. The multivariate dispersion of functional gene for each site was estimated using the Marti Anderson's PERMDISP2 procedure (Anderson 2006), which was based on the Euclidean distances between site samples and the site centroid on the principal coordinate axes. Beta diversity for each site was estimated using Whittaker's definition (Whittaker 1960) by

dividing gamma diversity (total functional gene richness) by alpha diversity (mean functional gene richness). Transformation of signal intensity to Z-score was performed for each probe across all samples, and then all probes were clustered using complete-linkage based hierarchical cluster analysis (Defays 1977) for signal intensity contrasting in a heat map. The significance of functional gene abundance differences between control and treatment samples was evaluated using the LSD test.

2.4 Results

2.4.1 Summary of PABMC probe design

The development of the PABMC started from ~130,000 amino acid sequences retrieved from NCBI protein database using manually crafted keyword query, and 42,605 of them were confirmed as valid targets for coverage of the PABMC. Further, about 20,000 coding sequences in NCBI nucleotide database were found to match up with confirmed protein targets and selected for probe design. More than 100,000 50-mer oligonucleotide probes were designed and then searched for specificity verification. Finally, a total of 3,870 best targeted probes were selected for synthesizing the PABMC in this study. These probes targeted 6,178 genes, coding sequences from 34 gene families, and are capable of detecting and identifying 1,761 PGPM species or strains, as listed in **Table 2.1**. Among these, 1,096 (28.3%) probes are gene-specific, targeting only a single gene sequence, while 2,774 (71.7%) probes are group-specific, targeting two or more gene sequences sharing a very high similarity among them. The PABMC also has both positive and negative controls for hybridization validation and data normalization, including 640 positive control probes (80 replicates × 8 degenerate probes) targeting 16S rRNA sequences, 1689 negative control probes specifically

targeting seven hyperthermophile genomes, and common oligonucleotide reference standard probes for data normalization and comparison.

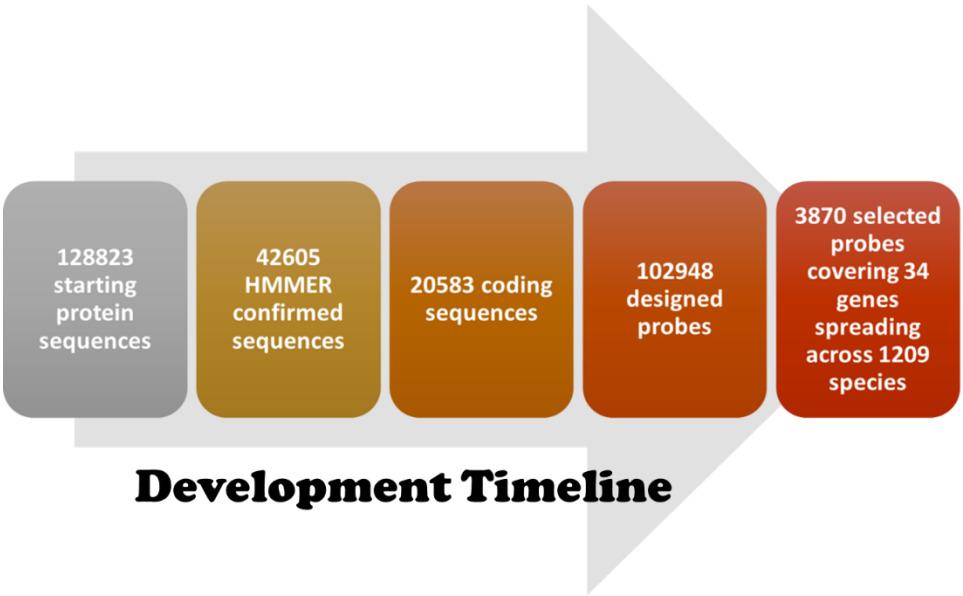


Figure 2.2 PABMC design results yielded from each intermediate step (timeline is from left to right). The PABMC development started from 123,823 candidate sequences, and only 42,605 sequences were confirmed to be sequences of interest using Hidden Markov Model screening. 20,583 nucleotide sequences were found on the basis of confirmed protein sequences, and 102,948 raw probes were designed. Finally, 3,870 probes have been selected for PABMC synthesis.

2.4.2 Selected functional genes for PABMC

Based on beneficial mechanisms, we divided selected genes into six functional categories: (i) pathogen resistance, (ii) promotion of plant hormone production, (iii) antibiotic synthesis, (iv) antioxidants, (v) drought resistance, and (vi) other beneficial processes. The rationale for selecting these functional groups is described below.

(i) Pathogen resistance genes. PGPMs can protect plants from disease or reduce their susceptibility. Siderophores are small, high-affinity iron chelating compounds generally produced under iron-limiting conditions to scavenge iron (Bossier, Hofte et al. 1988, Neilands 1995). The synthesis of siderophores in PGPMs provides an important mechanism to suppress pathogens (Miethke and Marahiel 2007). Thereby, a total of 299

probes were developed for siderophore biosynthesis protein (*sid*), including 187 group-specific probes and 112 gene-specific probes and covered 478 bacterial, archaeal and fungal siderophore biosynthesis protein coding sequences from 145 strains (**Table 2.1**).

Table 2.1 Summary of plant beneficial gene probes on the PABMC

Functional Category	No. of functional gene families	No. of group-specific probes	No. of sequence-specific probes	No. of covered CDS's	No. of covered microbial species (strains)
Pathogen resistance	3	187	112	478	145
Antibiotic	10	695	114	1092	459
Antioxidant	8	853	443	2640	430
Drought tolerance	2	119	233	624	85
Hormone promotion	8	725	138	1043	487
other	3	195	56	301	155
Total	34	2774	1096	6178	1761

(ii) Plant hormone biosynthesis genes. Soil microorganisms can release plant hormones to accelerate plant growth, stimulate germination and elongation, break dormancy, stimulate bolting, and delay senescence (Srivastava 2002, Osborne and McManus 2005). A total of 863 probes were developed for hormone production genes, including 725 group-specific probes and 138 gene-specific probes, covering 1043 gene sequences. Functional gene families chosen in this section include those coding gibberellin biosynthesis protein (*gas*), necrosis and ethylene-inducing protein (*nep*), ethylene biosynthesis protein (*eth*), spermine (*spe*) biosynthesis protein, cytokinins biosynthesis protein (*cks*), and spermidine synthase (*sped_bac* for bacterial, *sped_arc* for archaeal and *sped_fun* for fungi).

(iii) Antibiotic biosynthesis genes. Antibiotic biosynthesis by PGPMs serves as a competition strategy to protect themselves and suppress competitors (Raaijmakers and Mazzola 2012), from which plants may receive benefits if their antagonistic microbial

counterparts or pathogens were suppressed. A total of 809 probes were selected and designed for antibiotic activity related genes. These included 695 group-specific and 114 gene-specific probes, covering 1090 nucleotide gene sequences. Antibiotic biosynthesis-related proteins covered bacilysin biosynthesis protein (*bacA*), linear gramicidin biosynthesis protein (*lgrD*), lincomycin biosynthesis protein (*imbA*), chloramphenicol biosynthesis protein (*pabA*), isopenicillin N synthesis protein (*pcbC*), phenazine biosynthesis protein (*phzF*), epidermin biosynthesis protein (*epiA*), pyrrolnitrin biosynthesis protein (*prnB*), subtilin biosynthesis protein (*spaR*) and streptomycin biosynthesis protein (*strR*).

(iv) Antioxidant biosynthesis genes. Synthesizing antioxidant enzymes is a key defense mechanism for microorganisms to prevent reactive oxygen chemical species from producing hydroxyl radicals, thereby limit or prevent cell damage (Mates 2000). Antioxidant enzymes in soil can benefit plants not only by preventing oxidative injuries (Gianfreda 2015) to plant roots but also by protecting mutualistic plant-microbe symbiosis (Santos, Hérouart et al. 2000). Here, 1,296 probes were developed for antioxidant biosynthesis genes; 853 were group-specific and 443 were gene-specific. These probes were designed to cover 2,640 gene sequences encoding catalase (*cat_bac* for bacterial genes, *cat_arc* for archaea genes and *cat_fun* for fungal genes), peroxidase (*per_bac* for bacterial genes, *per_arc* for archaea genes and *per_fun* for fungal genes), and superoxide dismutase (*sod*).

(v) Drought resistance genes. Trehalose is a natural alpha-linked disaccharide that can be synthesized by microorganisms to work as a water retainer in soil (Luyckx and Baudouin 2011). When cells experience dehydration, trehalose can form a gel phase to

prevent disruption of internal organelles, assisting plants with prolonged desiccation tolerance (Luyckx and Baudouin 2011). This functional category includes a total of 352 selected probes; 233 were gene-specific and 119 were group-specific. These probes were designed to cover 624 gene sequences encoding trehalose synthase (*tre_arc* for archaea genes and *tre_fun* for fungal genes). However, no probes have been selected on the basis of our design criteria for covering bacterial strains.

(vi) Secondary beneficial genes. Three genes families that can bring indirect benefits to plants were included in this catalog, including those encoding pectinase (*pec*) and lipopolysaccharides biosynthesis protein (*lipo*) as elicitors of plant immune defense response, as well as 1-aminocyclopropane-1-carboxylate deaminase (*acsD*) as precursors of plant hormone. A total of 251 probes have been designed for covering 301 coding sequences of *pec*, *lipo* and *acsD*; 195 were group-specific and 56 were gene-specific.

2.4.3 Computational evaluation of specificity

The specificity of all designed probes was assessed computationally with respect to sequence identity, continuous stretch length and free energy (He, Wu et al. 2005, He, Deng et al. 2010, Tu, Yu et al. 2014). The maximum identity, maximum stretch length, and minimal free energy of each probe to their non-target sequences were measured. Approximately 88% of the probes had $\leq$ 60% maximum sequence identity to non-targets, and only 4% of probes fell within the range of > 86%. About 6% of probes had more than 18 bases of maximum continuous stretch (He, Wu et al. 2005) and others (94%) with 17 or few bases. Only 0.6% had -35 ~ -30 kcal mol^{-1} free energy to non-targets, and other (96%) with > -20 kcal mol^{-1} free energy to non-targets (**Figure 2.3**).

This assessment indicated most of the designed probes should have specific hybridization with their targets. Similarly, group-specific probe specificity was evaluated for their minimum sequence identity, minimum stretch length, and maximum free energy with their target sequences within a group. More than 97% of probes had a perfect identity to their targets, and only 2% of probes had stretch lengths shorter than 35 bases and 1% of probes had higher than -60 kcal mol^{-1} free energy (**Figure 2.3**). The results indicated that a vast majority of group-specific probes were very close to their targets in identity, stretch length and free energy. All results here showed the designed probes should be specific to their targets.

Table 2.2 Dissimilarity tests of soil microbial communities sampled from three (native, mixed and invaded) regions using three statistical methods based on all plant beneficial genes detected.

Sample	MRPP		ANOSIM		Adonis	
	δ	*P*	R	*P*	F	*P*
Among three groups	0.112	0.004	0.720	0.003	3.680	0.022
Invaded *vs.* Mixed region	0.118	0.103	0.666	0.114	3.387	0.010
Invaded *vs.* Native region	0.102	0.128	1.000	0.113	4.676	0.001
Mixed *vs.* Native region	0.118	0.101	0.629	0.087	3.288	0.192

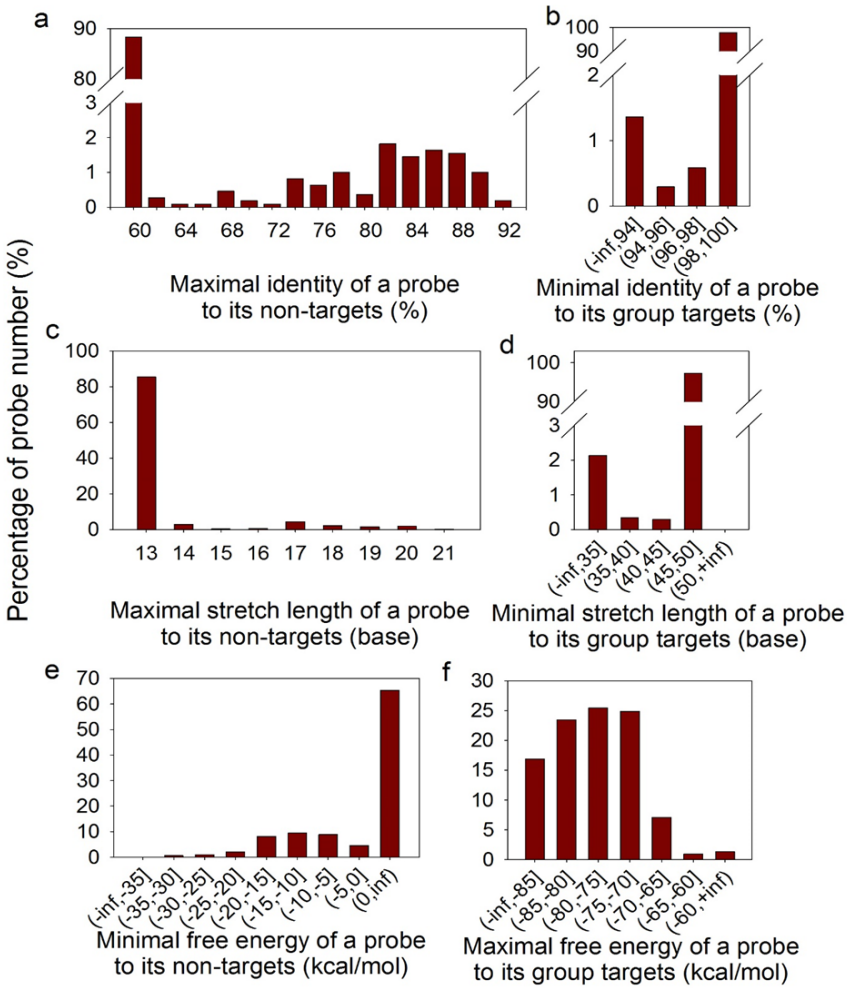


Figure 2.3 Computational evaluation assessed the specificity of all designed probes on the basis of sequence identity (a, b), stretch length (c, d) and free energy (e, f). Left panels (a, c, e) showed the assessment of sequence-specific probe specificity to the non-target sequences. Right panels (b, d, f) showed the assessment of group-specific probe specificity of deigned probes to their target sequences.

2.4.4 Application of the PABMC to characterize PGPM communities under exotic plant invasion

To understand how soil microbial communities respond to exotic plant invasion, we applied the developed the PABMC to analyze soil samples from native, mixed and invaded sites. A total of 1499 probes showed positive hybridization signals from the invaded site, 1521 probes from the mixed and 1411 probes from the native site. Plant

beneficial gene diversity was also disparate by the site. Three alpha diversity indices (gene richness, Shannon Index, and Simpson Index) were significantly higher in the invaded site than the native site, and no significant difference between mixed site and native site or between the mixed site and invaded site were found. The mixed site had greater alpha diversity variance (**Figure 2.4**) and beta diversity than other two sites (**Figure 2.5**). The NMDS analysis compared the overall composition and structure of plant beneficial gene in different sites. The results of NMDS provided a good fit of two-dimensional ordination (stress value = 0.03) on plant beneficial gene dissimilarities, and Shepard stress plot showed that the dissimilarities were strongly correlated ($R^2 = 0.94$) with the ordination distances (**Figure 2.6**). NMDS showed a clear separation by site, suggesting that the plant beneficial gene structure was altered by different intensities of exotic invasion (**Figure 2.6**). Three complementary non-parametric multivariate statistical tests revealed that the composition of soil microbial functional genes beneficial to plants differed significantly among three sites (MRPP: $\delta = 0.112$, $p = 0.001$, ANOSIM: $R = 0.072$, $p = 0.003$; ADONIS: $F = 3.680$, $p = 0.001$) or between any two pairs (**Table 2.2**).

Further analysis of detected probes was presented in a heat map, showing that different sites had distinctive probe signal intensity distribution (**Figure 2.7**a). Analysis of gene families (**Figure 2.7**b) found that the abundance of *tre_arc* from the category of drought tolerance in the mixed and invaded sites was significantly higher than that on the native site. Meanwhile, the abundance of *sid_arc* was significantly lower in the invaded site than in the native site. The mixed responses of gene abundance to *A. adenophora* invasion were observed for antibiotic synthesis genes and antioxidant biosynthesis

genes. Among them, two antibiotic synthesis genes (*lgrD* and *pabA*) significantly decreased in the invaded site, while another three genes (*lmbA*, *phzF* and *spaR*) significantly increased. Similarly, the abundance of two antioxidant biosynthesis genes (*cat_arc* and *sod_nickel*) increased and one gene (perl_bac) decreased significantly. Nevertheless, the abundances of genes involved in plant hormone biosynthesis and other genes (*pec*, *lipo* and *acsD*) didn't show any significant difference by the site. Overall, the PGPM community functional structure and some functional potentials were altered by the *A. adenophora* invasion, towards strengthened stress tolerance, but less pathogen resistance.

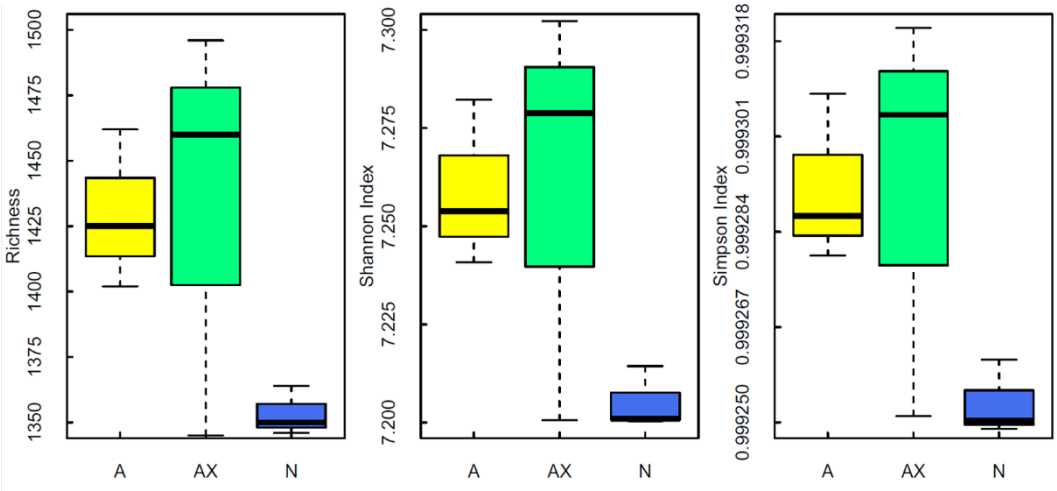


Figure 2.4 Diversity of plant beneficial genes in microbial communities in samples from the A (active; yellow), AX (mixed; green) and N (native; blue) site, calculated as functional genes richness, Shannon index, and Simpson index.

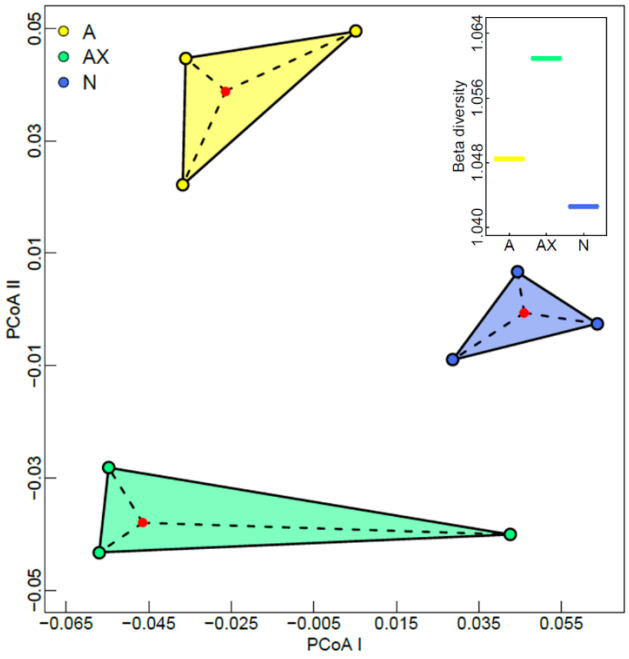


Figure 2.5 Multivariate dispersion and beta diversity of plant beneficial gene. Samples from the A (yellow), AX (green) and N (blue) site was plotted on two principal coordinate axes, and centroid for each site was positioned by red dots. The Euclidean distance between each sample and the corresponding centroid was plotted using dashed black lines. The inner plot indicated that beta-diversity for each site based on Whittaker's definition.

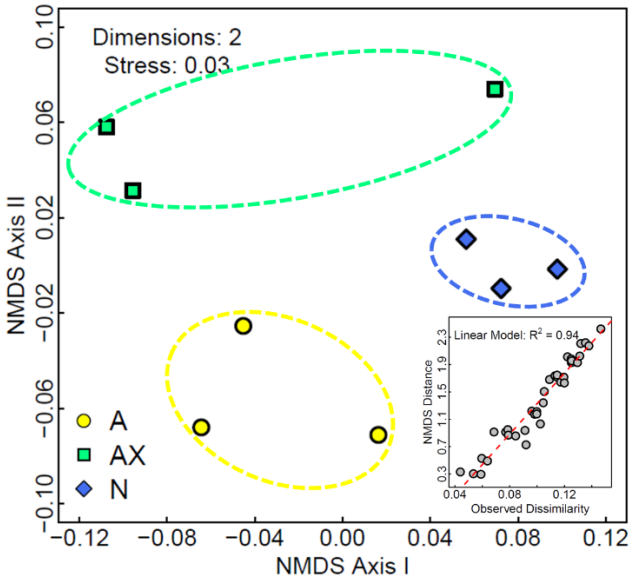


Figure 2.6 Non-metric Multidimensional scaling (NMDS) analysis of plant beneficial genes detected in A, *A. adenophora* invaded region; AX, *A. adenophora* and native plants mixed region, and N, native plants growing region.

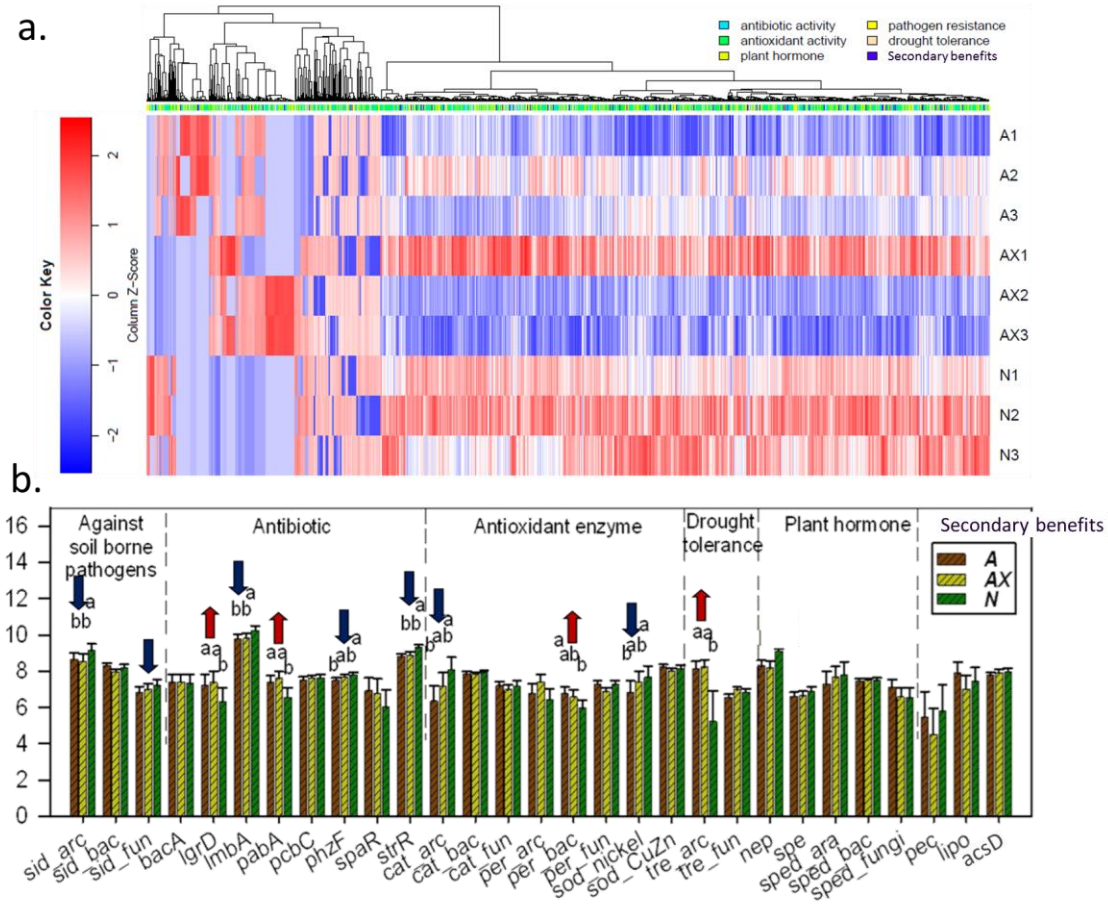


Figure 2.7 (a) Heat map of probe signal Z-score transformed from signal intensity across all samples. All probe signal Z-scores were clustered using complete-linkage based hierarchical cluster analysis for contrasting purpose. (b) Normalized relative abundances of plant beneficial genes detected by the PABMC. Antibiotic and stress tolerance genes increased their abundances in the invaded samples, while the abundance of pathogen repressing genes decreased.

2.5 Discussion

In this study, we developed the PABMC and used it to investigate changes of plant beneficial genes in response to plant invasion. To our knowledge, PABMC is the first high-throughput functional gene array to characterize plant beneficial genes with comprehensive coverage in terms of plant beneficial genes and PGPM species. This tool was firstly applied in studying PGPM responses to plant invasion. Detecting and profiling soil microbial functional genes beneficial to plants helps understand

interactions between PGPMs and plants in agricultural and natural ecosystems, which is of economic and ecological importance.

Comparing with traditional methods (e.g. culturing dependent techniques)(Coats and Rumpho 2014), the PABMC has higher throughput as it allows to simultaneously detect a broad range of hallmark genes involved in a variety of mechanisms providing plant beneficial effects. Thus, functional profile of relatively more comprehensive PGPM communities can be captured with single hybridization of the PABMC.

Other high-throughput platforms, like sequencing (either amplicon metagenomics sequencing or shotgun metagenomics sequencing) may be used to character PGPMs as well. However, 16S rRNA amplicon sequencing cannot discern PGPMs and obtains taxonomical or phylogenetical profiles for the whole soil microbial community, while functional gene amplicon sequencing might be restricted by a narrow inclusion of genetic markers that were insufficient for covering a majority of PGPM species. Also, detection of rare species or sequences in both methods is likely skewed by dominant species or contaminants, which might be a serious issue for profiling PGPMs which might be rare and unusual. Sensitivity and quantitativity issues for sequencing are mostly caused by selective PCR process that is a prerequisite for amplicon sequencing, but can be avoided for functional gene array. In contrast, functional gene arrays enable rapid and cost effective metagenomics comparative analysis across samples and ecosystems, which allows more investment in replication which is crucial for the reproducibility and success of metagenomics studies.

The functional gene array based technologies (e.g. GeoChip (He, Gentry et al. 2007, He, Deng et al. 2010, Tu, Yu et al. 2014), PathoChip (Lee, Van Nostrand et al. 2013)

and StressChip (Zhou, He et al. 2013)) have been shown in a number of studies to be effective in studying composition, structure, diversity and dynamics of microbial communities and establishing linkages between microbial communities and environmental variables in a variety of ecosystems (Zhou, Kang et al. 2008, Hazen, Dubinsky et al. 2010, He, Xu et al. 2010, Zhou, Xue et al. 2012, Zhou, Liu et al. 2013, Xue, M. Yuan et al. 2016). The PABMC was developed in a way consistent with the development of GeoChip, which was computationally and experimentally evaluated to be specific, sensitive and quantitative in detecting taxa and functional genes (He, Gentry et al. 2007, He, Deng et al. 2010, Lee, Van Nostrand et al. 2013, Zhou, He et al. 2013, Tu, Yu et al. 2014).

Specificity is one of the most critical issues in functional gene array technology, especially for characterizing microbial communities with complicated structures and composition. In PABMC development, probe specificity control was evaluated and verified using criteria that were proved to work in previous GeoChip versions (He, Wu et al. 2005, Liebich, Schadt et al. 2006). These criteria used for designing both gene-specific and group-specific probes have been well evaluated and established in a series of experiments (He, Wu et al. 2005, Liebich, Schadt et al. 2006). Testing results of previous studies showed that probes designed with these criteria were highly specific for sequence targeting (He, Wu et al. 2005, Liebich, Schadt et al. 2006). In addition, sequence overall similarity, continuous identical subsequence length and hybridization free energy were considered at the same time for high quality probe selection by CommOligo 2.0, ensuring all designed probes were specific to all coding sequences in the input file and with similar thermodynamic properties (Li, He et al. 2005).

Additionally, all designed probes were checked against the NCBI *nt* and *env_nt* databases for specificity, and non-specific probes were discarded. Fourth, computational evaluation showed that only a trivial portion of the designed probes (~1%) were close to the thresholds of previously established criteria (He, Wu et al. 2005). During the probe design, the total number of probes initially designed was more than 100,000, but only ~3.6% of them were used for PABMC synthesis after specificity verification and probe selection, and this should greatly reduce the risk of non-target cross hybridization. During preprocessing of hybridization results, the cutoff of probe intensity was set at 1,000 and the cutoff of SNR was set at 2, which should also reduce the false positives (He, Deng et al. 2010, Lu, He et al. 2012).

Exotic plant invasion causes regional aboveground biodiversity loss. To investigate the effectiveness of PABMC and responses of PGPMs to the exotic plant invasion, the PABMC was applied to characterize soil microbial communities from a region have been invaded by *A. adenophora*. It uncovered the impacts of plant invasion on microbial communities harboring plant beneficial genes, and offered metagenomic insights into how invasive plant interacted with and receive benefits from PGPMs and eventually established successional success. Composition and structure of plant beneficial gene in the invaded site is different from what in the native site, suggesting PGPM community was shifted and its diversity was increased by the *A. adenophora* invasion. The *A. adenophora* invasion also promoted the diversity of plant beneficial gene, likely for its own good as microbial diversity was closely tied with plant production and other terrestrial ecological functions (Wagg, Bender et al. 2014, Delgado-Baquerizo, Maestre et al. 2016). The higher microbial diversity was likely

caused by the plant-microbe interactions in soil, especially negative feedback (e.g. allelochemicals toxin production), which are the forces reorganizing microbial composition, and leading to species replacement and diversification (Reynolds, Packer et al. 2003). This speculation was supported by a recent finding of *A. adenophora* may release allelochemicals in root exudates (Yang, Qiu et al. 2013). Once a more diverse PGPM community occurs in soil, more diverse forms of benefits may be unlocked to *A. adenophora*, and led to its successful invasion.

Interestingly, the mixed site had greater the highest beta diversity than both the invaded and native site, suggesting PGPM community shift depends on the intensity of *A. adenophora* invasion. Because the mixed site could be seen as an intermediate state during the invasion, and PGPMs in the mixed site could be under the bilateral influences derived from both native plants and *A. adenophora*, different samples from the mixed site were possibly less homogenized as a result of the bilateral influences. This result is also consistent with a more general pattern that plant diversity promotes microbial beta diversity [50]. However, it may require data of multiple time points as snapshots for different invasion phases to confirm the implication.

The abundance of plant beneficial gene was changed by *A. adenophora* invasion in a way that could favor the establishment of *A. adenophora*. The potential for drought tolerance increased in the invaded site. Yunnan province is under the seasonally arid climate (dry season usually falls between November and April), and an extreme drought prolonged over the period of 2009-2010 (Yang, Gong et al. 2012). Thus, water preserving mechanism may be important for the success of the invasion. Our results revealed that *A. adenophora* may also gain from escaping the local pathogen

suppression as a strategy to gain fitness. The pathogen resistance potential in the invaded site was significantly lower than in the native site, which may support the "Enemy Escape Hypothesis" (Keane and Crawley 2002). This hypothesis suggests that *A. adenophora* was not affected by or was less affected by the native pathogens. In addition, there were both significant increases (2 genes) and decreases (3 genes) in the abundance of antibiotic biosynthesis genes. We speculated that *A. adenophora* might actively overturn belowground antibiotic regulatory landscape by suppressing original antibiotic synthesizers established in the native site in favor of substitute synthesizers. However, further studies are required to examine *A. adenophora* root exudates, soil properties, and the surrounding microbiota in order to confirm these hypotheses.

In conclusion, we developed the PABMC for detecting a broad range of plant beneficial genes from six categories, including plant growth-promoting hormones, plant pathogen resistance, antibiotics, antioxidants, drought tolerance, and secondary benefits (e.g. elicitor of plant immune defense response). We verified the specificity of the probes included in the PABMC was highly specific in the computational evaluation. In the showcase study to investigate PGPM communities in natural sites where have been invaded by *A. adenophora*, PGPM communities were shifted towards contributing to the success of *A. adenophora* invasion with increased diversity of genes benefitting plants, and changed relative abundance of genes in various categories, demonstrating PABMC as a powerful tool for characterizing the composition and structure of PGPM communities. It is also important to keep the PABMC updated for analyzing complex PGPM communities by incorporating more PGPM strains, and more plant beneficial

functional genes and categories in the future, if they will become state-of-art knowledge for PGPM study.

Chapter 3: Ultra-sensitive and -quantitative Detection of Microbial Populations in complex communities with New Functional Gene Arrays

3.1 Abstract

The rapid development of high-throughput metagenomic technologies over the past decade has greatly extended our understanding of complex microbial systems. While remarkable advances have been made in the development of high-throughput functional gene arrays (FGA) for analyzing complex microbial communities, challenges still remain in their representation, specificity, sensitivity, and quantitation. Here we developed a new generation of high-density FGA, GeoChip 5.0 based on Agilent platform, with two formats. The smaller format contained 60K probes (GeoChip 5.0S), majorly covering probes from carbon (C), nitrogen (N), sulfur (S), and phosphorus (P) cycling and energy metabolism probes. The larger format (GeoChip 5.0M) contained all probes in GeoChip 5.0S and expanded to antibiotic resistance, metal resistance/reduction, organic contaminant remediation, stress responses, pathogenesis, soil beneficial microbes, soil pathogens, and virulence. GeoChip 5.0M contains 161,961 probes covering approximately 370,000 representative coding sequences from 1,447 functional gene families. These genes were derived from functionally divergent broad taxonomic groups, including bacteria (2,721 genera), archaea (101 genera), fungi (297 genera), protists (219), and viruses (167 genera, mainly phages). Both computational and experimental evaluation with perfect match (PM)/mismatch (MM) probes indicated that all designed probes were highly specific to their corresponding targets. Good hybridization could be obtained with 100 ng DNA. Sensitivity tests revealed that as

little as 0.05 ng of pure culture DNAs was detectable within 1 μg of complex soil community DNA as background. This is equivalent to 0.005% of a population within a complex community, suggesting that the Agilent platform-based GeoChip is extremely sensitive. Additionally, very strong quantitative linear relationships were obtained between signal intensity and pure genomic DNAs (about 99% of probes detected with $r > 0.9$) or soil DNAs (about 97% of the probes detected with $r > 0.9$) within at least three orders of magnitudes. Application of the designed FGAs to a contaminated groundwater with very low biomass indicated that environmental contaminants (majorly, heavy metals) had significant impacts on the biodiversity of microbial communities. The GeoChip 5.0 developed in this study is the most comprehensive FGA directly linking microbial genes/populations to ecosystem processes and functions.

3.2 Introduction

Microorganisms are the most diverse and ubiquitous life on earth. They interact each other to form communities integral to various ecosystem processes and functions that are of critical importance in global biogeochemical cycling, human health, energy, climate change, environmental remediation, engineering, industry, and agriculture (Curtis, Head et al. 2003, Zhou, Deng et al. 2014). Despite their importance, however, determining microbial community structure and functions remains challenging for several reasons. First, microbial diversity is extremely high. Studies indicated that one gram of soil could contain 2,000 to 8.3 million species (Gans, Wolinsky et al. 2005, Schloss and Handelsman 2006, Roesch, Fulthorpe et al. 2007), a majority of which (>99%) have not been cultivated (Rappe and Giovannoni 2003). Numbers of microbial cells from environmental habitats are also extremely large, thus it is also impossible to directly count the cells. For example, such a number was estimated to be 1.2×10^{29} in the open ocean (Whitman, Coleman et al. 1998), 2.9×10^{29} in the sub-seafloor sediment (Kallmeyer, Pockalny et al. 2012), and 2.6×10^{29} in soil (Whitman, Coleman et al. 1998). These communities also represent a high diversity of functional potential (Sogin, Morrison et al. 2006). Establishing mechanistic linkages between microbial biodiversity and ecosystem functioning poses another grand challenge for microbiome research. To tackle these challenges, more advanced high-throughput metagenomics technologies for characterizing complex microbial communities are needed (Zhou, He et al. 2015). Most recently, several types of high-throughput technologies have been developed to characterize microbial communities, including next generation sequencing (Venter, Remington et al. 2004, Frias-Lopez, Shi et al. 2008, Caporaso, Lauber et al. 2012,

Loman, Misra et al. 2012, Weinstock 2012), microarrays (e.g., PhyloChip , GeoChip (He, Deng et al. 2010)), quantitative PCR (Arya, Shergill et al. 2005), mass spectrometry-based proteomics (Ram, VerBerkmoes et al. 2005), and metabolomics (de Raad, Fischer et al. 2016). These technologies have provided unprecedented insights into our understanding of microbial biodiversity and detection of novel processes and functions (Valdes, Glass et al. 2013). Among these, high-throughput sequencing and microarrays are two of the most widely used open and closed format technologies (Zhou, He et al. 2015), with distinct differences in susceptibility to random sampling errors and non-targeted DNAs, ability to detect novel organisms and rare species, capability of quantitation, and difficulties in data analysis (Zhou, He et al. 2015). Consequently, both have unique advantages and disadvantages in detection specificity, sensitivity, quantification, resolutions, and reproducibility (Zhou, He et al. 2015). It is highly beneficial if both types of technologies are used in complementary fashions to address fundamental questions in microbial ecology (Zhou, He et al. 2015).

Over the last few decades, a variety of DNA microarray-based technologies have been developed for microbial detection and community analysis (He, Van Nostrand et al. 2011), such as phylogenetic and functional gene arrays (Zhou 2003). Phylogenetic gene arrays often contain probes from phylogenetic markers such as rRNA genes, which are useful for taxonomical profiling in microbial communities and investigating phylogenetic structures. Functional gene arrays (FGAs) target genes involved in various functional processes (Zhou, He et al. 2015), which are valuable for assessing the functional composition and structure of microbial communities. Although various types of FGAs are available (Zhou, He et al. 2015), GeoChip, a generic FGA targeting

hundreds of functional gene categories important to biogeochemical, ecological, and environmental analyses, is mostly widely used. GeoChip has been shown to be an effective, sensitive and quantitative tool for examining the functional structure of microbial communities (Wu, Liu et al. 2006, Brodie, DeSantis et al. 2007, Zhou, Kang et al. 2008, Hazen, Dubinsky et al. 2010, He, Xu et al. 2010, Zhou, Xue et al. 2012) from a variety of environments (He, Deng et al. 2012, Trivedi, He et al. 2012), including soils (Zhou, Kang et al. 2008, He, Xu et al. 2010, Trivedi, He et al. 2012, Yergeau, Bokhorst et al. 2012, Zhou, Xue et al. 2012), aquatic ecosystems (Taş, van Eekert et al. 2009, Kimes, Van Nostrand et al. 2010), extreme environments (Wang, Zhou et al. 2009, Mason, Nakagawa et al. 2010), contaminated habitats (Leigh, Pellizari et al. 2007, Liang, Li et al. 2009, Liebich, Wachtmeister et al. 2009, Van Nostrand, Wu et al. 2009, Xiong, Wu et al. 2010, Xu, Wu et al. 2010, Liang, Van Nostrand et al. 2011) and bioreactors (Liu, Wang et al. 2010, Liu, Zhang et al. 2012).

Although many technical issues regarding microarray technology have been solved, several critical bottlenecks still exist. One of the greatest challenges is that most of the probes on the current GeoChip were derived from genes/sequences available in publicly available databases and do not necessarily fully represent the diversity of the microbial communities of interest given the rapid expansion of sequence information in public databases. Consequently, it could be difficult to use the current GeoChip to fully address research questions in a comprehensive manner if the gene probes on the array do not represent the diversity of the microbial communities examined. Thus, further developments are needed to improve its representativeness and performance in terms of specificity, sensitivity, and quantitation. In this study, we aimed to develop a new

generation of more comprehensive and representative FGA, termed GeoChip 5.0. All previous functional gene families have been updated and more than 1,000 new functional gene families have been added, including those involved in metal homeostasis, secondary metabolism, virulence, and phylogenetic markers for fungi, protists, and viruses. The newly developed GeoChip 5.0 was systematically evaluated in terms of specificity, sensitivity, and quantitative capability. It was then applied to analyze the responses of groundwater microbial communities to high concentrations of U(VI) and nitrate as well as low pH. Our results demonstrate that the developed GeoChip is highly specific, sensitive, and quantitative for functionally profiling microbial communities.

3.3 Materials and methods

3.3.1 Sequence retrieval and probe design

Sequence retrieval and probe design for GeoChip 5.0 were performed using the GeoChip design pipeline as described previously (He, Deng et al. 2010, Tu, Yu et al. 2014). Briefly, a keyword query for each protein-encoding gene was submitted to the NCBI nr database to retrieve candidate sequences (**Figure S 2**). Next, sequences that had been experimentally confirmed for each protein/enzyme were selected as seed sequences, which were then used for building a Hidden Markov Model (HMM) to search homologs against and confirm each candidate sequence. Confirmed sequences were potential targets for probe design. Then all the targets were searched against the legacy probes from previous versions of GeoChip. This was done to determine if any targets were covered by legacy probes or if any legacy probes were no longer valid. All targets covered by legacy probes were directly assigned to the corresponding probe and

excluded from further probe design. Those legacy probes that were no longer valid were removed from the collection and the corresponding targets were released and reused for probe design. The probe design for novel targets (e.g. targets there were not previously covered by any legacy probes) and released targets were performed using a new version of CommOligo (Li, He et al. 2005). Two types of probes were designed: gene-specific (each probe targets one gene sequence); and group-specific (one probe targets two or more highly homologous sequences) (He, Deng et al. 2010). Finally, the newly designed candidate probes and all probes from previous GeoChip versions were searched against the NCBI nt/env_nt databases to verify their specificity.

3.3.2 Microarray construction

Two major formats of the GeoChip 5.0 array were developed. The smaller format (GeoChip 5.0S) has ~60,000 probes per array. For testing various experimental parameters, various modifications of GeoChip 5.0S also were made by including various perfect match (PM) and mismatch (MM) probes from different pure cultures. The larger format (GeoChip 5.0M) has ~180,000 probes per array. All GeoChip 5.0 microarrays were manufactured by Agilent (Santa Clara, CA, USA) using either the 8 x 60 K (8 arrays per slide) or the 4 x 180 K (4 arrays per slide).

3.3.3 DNA extraction, purification, and quantification.

Genomic DNA from *Desulfovibrio vulgaris* and *Clostridium cellulolyticum* were extracted using a GenElute Bacterial Genomic DNA Kit (Sigma-Aldrich) in accordance with the manufacturer's instructions and recommended pretreatment for Gram-negative bacteria. Microbial community samples used to evaluate GeoChip 5.0 performance were obtained from BioCON experimental site (Reich, Knops et al. 2001). The soil

microbial community DNA was extracted by freeze-grinding mechanical lysis (Zhou, Bruns et al. 1996) and purified using a low-melting agarose gel followed by phenol extraction. Groundwater samples from the Oak Ridge Integrated Field Research Center (Smith, Rocha et al. 2015) were used to evaluate the applicability of newly developed GeoChip.

DNA quality was assessed based on absorbance ratios (A260/A280 and A260/A230) using a NanoDrop ND-1000 spectrophotometer (NanoDrop Technologies Inc., Wilmington, DE), and DNA concentrations were measured with PicoGreen (Ahn, Costa et al. 1996) using a FLUOstar Optima microplate reader (BMG Labtech, Jena, Germany).

3.3.4 Target DNA preparation, amplification and labeling

Earlier versions of GeoChip used 1,000 ng of DNA for hybridization, so this amount was used as a starting for point for the 5.0M version and 500 ng DNA was used for 5.0S, since it is half the size of the larger version. The optimal DNA concentrations for hybridization were determined with different amounts of DNA templates, ranging from 1 ng to 1000 ng.

Whole community genome amplification (WCGA) was used to increase the available DNA if there was not enough DNA available. Aliquots (5-10 ng for groundwater samples) of DNA were amplified using the Templiphi kit (GE Healthcare) and a modified reaction buffer containing 0.1 mM spermidine and 267 ng ml^{-1} single stranded binding protein to improve the amplification efficiency (Wu, Liu et al. 2006). Samples were amplified for 6 h. For those groundwater samples without measurable DNA (by PicoGreen), the samples were concentrated to a volume of 10 µl and 2-5 µl was used for

amplification (initial amplification was attempted with 5 µl and then reduced if unsuccessful). All amplified products (~2 µg) was used for labeling and hybridization. DNA (amplified or unamplified) was mixed with 5.5 µl random primers (Life Technologies, random hexamers, 3 µg/µl), brought to 35 µl with nuclease-free water, heated to 99 °C for 5 min, and immediately placed on ice. Labelling master mix (15 µl), including 2.5 µl of dNTP (5 mM dAGC-TP, 2.5 mM dTTP), 0.5 µl of Cy-3 dUTP (25 nM; GE Healthcare), 1 µl of Klenow (imer; San Diego, CA; 40 U ml^{-1}), 5 µl Klenow buffer, and 2.5 µl of water, was added and the samples were incubated at 37°C for 6 h in a thermocycler and then at 95°C for 3 min to inactivate the enzyme. After the addition of Cy3, samples were protected from the light as much as possible. Labeled DNA was cleaned using a QIAquick purification kit (Qiagen) per the manufacturer's instructions and then dried down in a SpeedVac (45°C, 45 min; ThermoSavant).

3.3.5 GeoChip hybridization

Because there are two versions of GeoChip 5.0, GeoChip 5.0S and 5.0M, each version uses a different volume of hybridization buffer. The volumes below are the standard hybridization conditions for the GeoChip 5.0M, volumes for the GeoChip 5.0S are in parentheses.

Labeled DNA was resuspended into 27.5 µl (11.9 µl) of DNase/RNase-free distilled water, and then mixed completely with 99.4 µl (43.1 µl) of hybridization solution containing 63.5 µl (27.5 µl) of 2×HI-RPM hybridization buffer, 12.7 µl (5.5 µl) of 10×aCGH blocking agent, formamide (10% final concentration), 0.05 µg/µl Cot-1 DNA, and 10 pM common oligonucleotide reference standard (Liang et al., 2010). The solution was denatured at 95 ^{o}C for 3 min, and then incubated at 37 ^{o}C for 30 min. The

DNA solution was centrifuged briefly (1 min, 6000 x g) to collect liquid at bottom of tube and then 110 µl (48 µl) of the solution was pipetted into the center of the well of the gasket slide (Agilent). The array slide is placed on the gasket slide, array side down, sealed using a SureHyb chamber and then placed into the hybridization oven. The arrays were hybridized at 67 ^{o}C for 24 h.

After hybridization, slides were disassembled in room temperature Wash Buffer 1 (Agilent), then transferred to fresh room temperature Wash Buffer 1 on a magnetic stir plate set at 200 rpm and incubated for 5 min. Next, the slides were incubated at 37°C in Wash Buffer 2 (Agilent) for 1 min on a magnetic stir plate set at 140 rpm. Slides were then slowly removed from the buffer. The slide's hydrophobic coating allowed the slide to shed the buffer and dry almost immediately.

3.3.6 Microarray imaging and signal processing

The slides were imaged as a Multi-TIFF with a NimbleGen MS200 Microarray Scanner (Roche NimbleGen, Inc., Madison, WI, USA) and the data was extracted using the Agilent Feature Extraction program, v 11.5. Extracted data was then loaded onto the GeoChip data analysis pipeline (http://www.ou.edu/ieg/tools/data-analysis-pipeline.html).

Probe quality was assessed and poor or low signal probes were removed. Probe spots with the coefficient of variance (CV; probe signal SD/signal) > 0.8 were removed. Then the signal-to-noise ratio (SNR; (probe signal-background)/background SD) was calculated. In general, a local background that represents the actual background signal for each spot is used for calculations. As suggested by Agilent, we used the average signal of Agilent negative control probes in each sub-array as background signal for all

probes in the same sub-array instead of the local background. If all of the Agilent negative control probes within a given sub-array fail to yield a valid signal, then the mean background signal intensity from one of the adjacent sub-arrays would be used instead. The signal intensity for each spot was calculated for subsequent analysis by subtracting the signal intensity of Agilent negative spots within a sub-array. If the net difference is less than 0, the spots are excluded for subsequent analysis.

Data normalization and quality filtering were performed with two steps (Liang, He et al. 2010, Tu, Yu et al. 2014). First, the mean value of the signal intensity of the common oligonucleotide reference standard probes (CORS) (Liang, He et al. 2010) was calculated for each array, and then the signal intensity of samples was normalized using the maximum average value. Second, we calculated the sum of the signal intensity for each sample, and normalized the signal intensity of all spots in an array using the maximum sum value. A detailed description of the optimized GeoChip sample preparation, hybridization, imaging and normalization methods and reagents and equipment needed is in (Van Nostrand, Yin et al. 2016).

3.3.7 Statistical analysis

We used various statistical methods for GeoChip data analysis. Pearson correlation coefficient (r) was used for estimating linear relationships involved in this study. Three different nonparametric multivariate analysis methods, ADONIS (permutational multivariate analysis of variance using distance matrices), ANOSIM (analysis of similarities) and MRPP (multiresponse permutation procedure), as well as detrended correspondence analysis (DCA), were used to measure the overall differences of community functional gene structure (Zhou, Xue et al. 2012). The functional gene

diversity of microbial community was estimated using Shannon Index, Simpson Index, and functional gene richness. Pearson correlation coefficient was used for testing the dependence among the environmental factors, and hierarchical cluster analysis (hclust in R) was performed for identifying factor clusters. Canonical correspondence analysis (CCA) was used for analyzing the linkages between the functional gene structure and environmental factors. Welch's *t*-test was used for testing level of significance of the difference in functional gene richness and alpha diversity between paired groups of samples without assuming unequal variances.

3.4 Results

3.4.1 Selection of gene families and categories for array fabrications

All functional gene families from previous GeoChips (410) were updated and included in GeoChip 5. During this update, some gene families were combined or separated based on newly discovered gene families or increased sequence availability. For example, twelve dioxygenase gene families from GeoChip 4 were combined into three gene families due to similarities in the sequences of these families; *norB* was spilt into two gene families to differentiate a new subgroup discovered after the design of GeoChip 4. GeoChip 5.0 also greatly expanded overall gene and sequence coverage by adding more than 1,000 new gene families from functionally divergent broad taxonomic groups of bacteria, archaea, fungi, algae, protists, and viruses. The rationales for selecting various gene families were provided in various previous publications (He et al. 2007; 2009; Tu et al. 2014; (Lee, Van Nostrand et al. 2013), (Zhou, He et al. 2013, Van Nostrand, Zhou et al. 2016) (He, Gentry et al. 2007, He, Deng et al. 2010, Lee, Van Nostrand et al. 2013, Zhou, He et al. 2013, Tu, Yu et al. 2014).

GeoChip 5.0M covered a total of 1,447 gene families involved in carbon (135), nitrogen (28), sulfur (27), and phosphorus (7) cycling, antibiotic resistance (19), stress response (103), microbial defense (65), metabolic pathways (4), plant growth promotion (115), virulence (605), metal homeostasis (119), organic contaminant degradation (105), pigments (30) and electron transfer (11) (**Table 3.1**). The numbers of probes on GeoChip 5.0M were substantially more than those in GoeChip 4 for most of the functional gene categories, ranging from +29% to 368%. However, the numbers of probes in two gene categories (N cycling and organic contaminant degradation) decreased slightly due to more coverage by group-specific probes. From the taxonomic/phylogenetic perspective, GeoChip 5.0M had targeting probes from ~6,500 bacterial strains (2721 genera), 282 archaeal strains (101 genera), 625 fungi (297), 362 protists (219), 86 other lower eukaryotes (64), 1,364 viral strains (167 genera), and uncultured/unidentified organisms (**Table 3.2**). Compared to those in GeoChip 4, phylogenetic coverages in GeoChip 5 were substantially increased from 93% to 166%. Detailed information with respect to functional gene categories and phylogenies and their differences between GeoChip 4 and 5 were presented in Table S1 and S2.

Table 3.1 Summary of probes in GeoChip 5.0M based on functional gene categories.

Functional gene categories	No. of subcategories	No. of genes or enzymes	No. of sequence-specific probes	No. of group-specific probes	No. of total probes	No. of covered CDS	% probe changes since GeoChip 4
C Cycling	4	135	5,387	19,877	25,264	52,234	129%
N Cycling	9	28	2,572	3,739	6,311	12,179	-15%
S Cycling	8	27	2,229	2,510	4,739	7,439	52%
P Cycling	4	7	960	2,300	3,260	6,245	143%
Organic Contaminant Degradation	9	105	2,313	9,288	11,601	28,104	-32%
Antibiotic resistance	3	19	2,203	13,650	15,853	35,384	375%
Stress	24	103	2,792	25,038	27,830	82,183	29%
Metal Homeostasis	24	119	5,529	37,877	43,406	93,092	368%
Microbial Defense	3	65	1,321	6,426	7,747	14,989	NA
Metabolic Pathways	1	4	206	1,002	1,208	2,377	NA
Plant Growth Promotion	4	14	225	515	740	1,114	NA
Pigments	6	30	298	1,014	1,312	2,357	NA
Electron transfer	1	11	363	423	786	1,218	NA
Virulence	41	605	1,692	3,826	5,518	11,720	48%
Virus	3	115	1,521	1,336	2,857	5,182	167%
GyrB	1	1	532	2,234	2,766	9,997	NA
Protist	13	59	497	266	763	1,077	NA
Total	158	1,447	30,640	131,321	161,961	366,891	97%

*NA, not applicable because these are new addition for GeoChip 5.0

Table 3.2 Summary of probes in GeoChip 5.0M based on broad microbial groups.

Major microbial groups	No. of phyla	No. of genera	No. of species	No. of strains	No. of genes	No. of probes	No. of covered CDS	% probe changes since GeoChip 4
Bacteria	33	1122	2721	6465	1,003	141,153	333,675	93%
Archaea	6	101	188	282	269	5,728	38,978	124%
Fungi	7	297	404	625	226	8,856	21,101	130%
Protists	10	219	251	362	201	2,051	5,376	
Other eukaryota*	7	64	66	86	62	509	1,170	
Viruses	1	167	311	1,364	116	2,848	6,028	166%
Unclassified	-	-	-	-	125	816	2,561	116%
Total	64	1,970	3,941	9,184	1,447	161,961	366,891	97%

* Other eukaryota include Metazoa and Viridiplantae.

3.4.2 GeoChip 5.0 design and overall features

GeoChip 5.0 was *in situ* synthesized by Agilent with SurePrint technology. The spots are circular and are 30 microns in diameter. Compared to other array technologies, Agilent arrays have a wider dynamic range, higher sensitivity, and better quantitative capability. GeoChip 5.0S contained ~42K probes for ~95K target genes is focused on the analysis of key ecological and geochemical processes by covering only the core biogeochemical cycles (C, N, S, and P), and several important genes from other categories such as major facilitator superfamily antibiotics efflux pump genes (*MFS_antibiotic*), multidrug efflux transporter genes (*Mex*), nickel ABC transporter genes (*nikA*) and magnesium transporting ATPase (*mgtA*), degradation genes for relatively common contaminants (such as BTEX), and metals. GeoChip 5.0M is a more comprehensive design and contained ~162K probes from ~366K target genes, which covered all of the functions on the smaller array, also included a wider range of genes from additional functional categories across different organism groups (bacteria, archaea, fungi, algae, protists and viruses, **Table 3.1** and **Table 3.2**). GeoChip 5.0M was designed for a general survey of environmental, ecological and biogeochemical processes. A variety of probes were designed as controls for synthesis, hybridization, gridding and data analysis in both GeoChip 5.0S and GeoChip 5.0M (**Table 3.3**). For instance, GeoChip 5.0M contained 12,144 (~7%) probes that served as functional features for microarray synthesis, quality control, and position. A total of 4,096 degenerate probes targeting 16S rRNA sequences and 3,390 Agilent negative control probes served as positive and negative controls, respectively, for hybridization. To assist with normalization of signal intensity, GeoChip 5.0M had 3,378 probes targeting

seven sequenced hyperthermophile genomes and 1,280 common oligonucleotide reference standard probes (CORS) (Liang, He et al. 2010). The GeoChip array was arbitrarily divided into 256 (8 × 32; 5.0S) or 2,048 (8 × 256; 5.0M) grids. Control probes were placed so that each grid had 16 16S control probes and 5 CORS probes at specific positions. 16S control probes were splitted into two groups of 8, and were placed on each grid at the beginning of the first row and the end of the last row, respectively. CORS probes were placed on the central region of each grid. Each grid also had 2 or 3 Agilent negative control probes whose positions were randomized. The hyperthermophile and functional gene probes were randomly placed across the entire array in the available spot space.

Table 3.3 Summary of probes in GeoChip 5.0M based on broad microbial groups.

Entry	*GeoChip 4*	*GeoChip 5.0S*	*GeoChip 5.0M*	*% increases in GeoChip 5.0M since GeoChip 4*
Manufacturers	NimbleGen	Agilent		-
Feature shape	Square	Circular		-
Feature size	13 × 13 μm	30 micron (diameter)		-
Maximum features per array	135,000	60,000	180,000	-
No. of arrays per slide	12	8	4	-
No. of genes	410	308	1,447	+253%
No. of probes	82,074	41,781	161,961	+97%
No. of sequence-specific probes	18,098	10,252	30,640	+69%
No. of group-specific probes	63,976	31,529	131,321	+105%
No. of covered CDS	141,995	94,829	365,651	+158%
No. of covered strains	5247	4,859	9,195	+75%
16S positive controls	640	1,536	4,096	+540%
Controls from thermophiles	1,689	1,126	3,378	+100%
Universal standards	6,000	480	1,280	-78.6%
Agilent negative controls	-	1,565	3,390	-

3.4.3 Optimization of hybridization conditions

Hybridization for Agilent array with 60-mer probes is generally carried out at 65 ˚C to achieve good specificity with pure genomic DNAs (Barrett, Scheffer et al. 2004). However, GeoChip only uses 50-mer probes and is for detecting microbial populations in complex communities of unknown backgrounds, so hybridization conditions need to be optimized in terms of hybridization temperature, formamide concentration, and DNA amounts to achieve efficient and specific hybridization. First, the temperature is most important to determine hybridization specificity and efficiency. Different hybridization temperatures, ranging from 60 to 75 ˚C were tested (data not shown). Our results indicated that good hybridization can be achieved at 67 ˚C as judged visually (**Figure S 3**). Also, although the standard Agilent hybridization protocol does not use formamide, our previous studies indicated that adding formamide into hybridization buffer is useful to achieve high-specific hybridization with low background hybridization for environmental DNAs (Wu, Thompson et al. 2001, Rhee, Liu et al. 2004, He, Gentry et al. 2007, He, Deng et al. 2010, Tu, Yu et al. 2014). Thus, different formamide concentrations (0%, 10%, 15%, 20% and 25%) were evaluated. Our results suggested that efficient hybridizations at 67 ˚C with 10% formamide were obtained (**Figure S 3**). Template DNA concentration also has significant impacts on hybridization efficiency. Thus, different amounts of community DNAs were directly labeled with fluorescent dyes and hybridized with GeoChip 5.0S or 5.0M. Although the number of spots detected increased as DNA concentration increased, a good percentage of spots (> 30%) were obtained at 500 ng and 250 ng for both GeoChip 5.0S and 5.0M (**Figure S 4**a, b) respectively. About 18% of spots were detected even at 100 ng community DNAs for

GeoChip 5.0M. Based on above the results, 500 ng or 1000 ng community DNAs were generally recommended for hybridization at 67 ˚C plus 10% formamide as our standard hybridization conditions. 100 ng for direct labeling is acceptable for GeoChip 5.0S if the DNA concentration is really low.

3.4.4 Specificity of designed arrays

To determine if all designed probes are specific to their corresponding targets, we first computationally evaluated the probe specificity against our three design criteria (e.g., sequence identity of ≤ 90%, continuous stretch length ≤ 20 bases, and free energy ≥ -35 cal/mol). For sequence-specific probes, the maximum identity, maximum stretch length and minimal free energy to their closest non-target sequences were calculated. The majority of the designed sequence- or group-specific probes (82.2%) had less than 60% of maximum sequence identities to their non-target sequences in the NCBI databases (nt and env_nt) (**Figure 3.1**a). Less than 1% of the designed probes showed 86–90% sequence identity with their non-target sequences in the databases, and no probes had >90% sequence identity with their non-target sequences (**Figure 3.1**a). Also, the majority of the designed probes (93.8%) had maximal continuous sequence stretches of less than 19 bp to their non-target sequences in the databases (**Figure 3.1**c). In addition, about 99.3% probes had minimal free energy larger than -30 kcal/mol (**Figure 3.1**e). As previously demonstrated experimentally, the designed probes would be highly specific (Liebich, Schadt et al. 2006) if they have < 90-92% sequence identity, < 20 bp continuous sequence stretch, and > -35 kcal/mol free energy to their non-target sequences in the databases.

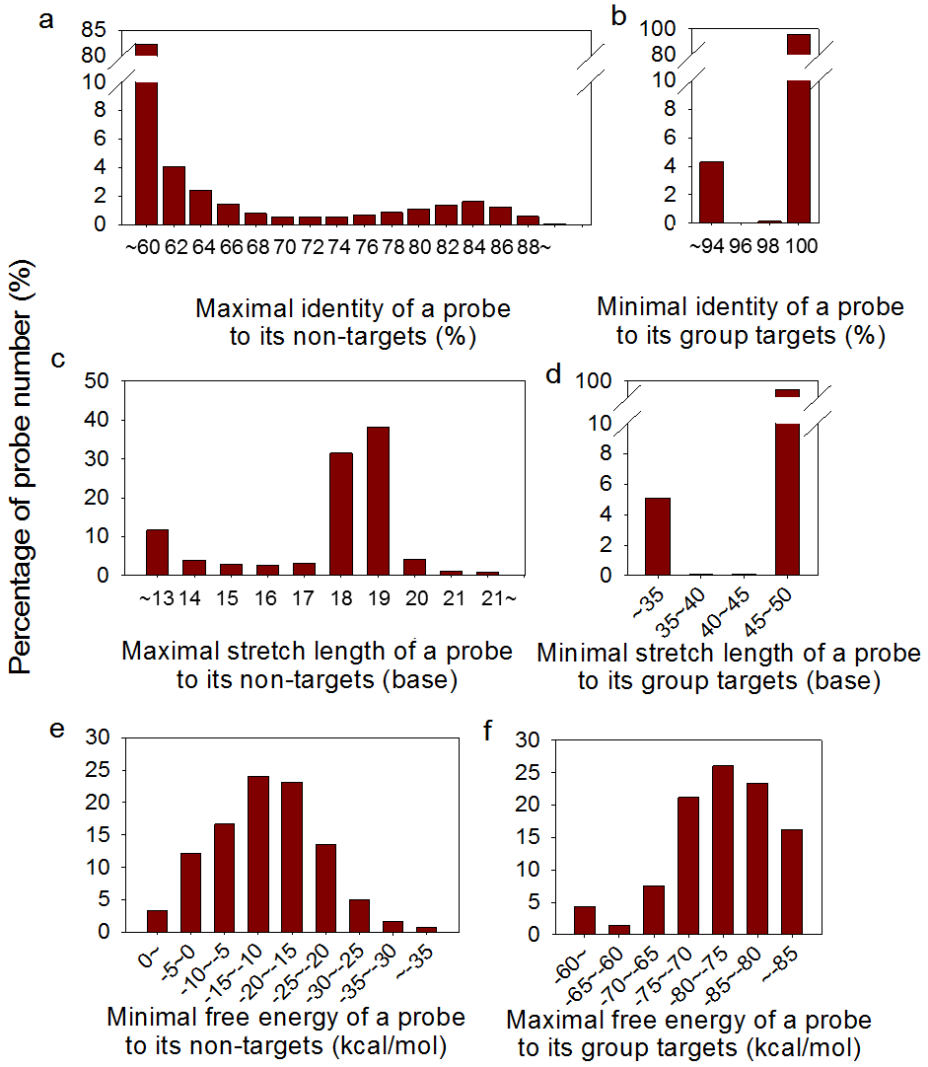


Figure 3.1 Computational evaluation of the specificity of the designed probes based on sequence identity, length of contentious sequence stretch and free energy. The three parameters were evaluated by comparing the designed probes to the sequences in the databases. (a) Maximal sequence identity (%) of a probe (sequence- or group-specific) to its closest non-target sequences. (b) Maximal sequence stretch length (bp) of a probe to its closest non-target sequences. (c) Minimal free energy (kcal/mol) of a probe to its closest non-target sequences. (d) Minimal sequence identity (%) of a group-specific probe to its targeted group sequences; (e) Minimal sequence stretch length (bp) of a group-specific probe to its targeted group sequences; and (f) Maximal free energy (kcal/mol) of a group-specific probe to its targeted group sequences.

For group-specific probes, there are potential mismatches between a group-specific probe and the corresponding target sequences. Such mismatches could affect the hybridization efficiency and hence the subsequent sensitivity and quantification. Thus,

we further required that the group-specific probes to have a minimal sequence identity of >94%, minimal continuous stretch length of > 35bp, and maximal free energy of < -60 kcal/mol (He, Deng et al. 2010, Tu, Yu et al. 2014). As shown in **Figure 3.1**b, d, and f, more than 94% of the designed group-specific probes had a sequence identity of ≥ 98%, continuous sequence stretch of ≥ 45 bp, and free energy of ≤ -70 kcal/mol to their corresponding target sequences. All of the above results were consistent with the probe design criteria (He, Deng et al. 2010, Tu, Yu et al. 2014), showing that the designed probes were highly specific to their target sequences and efficient hybridization with their target sequences can be achieved under the optimal experimental conditions.

The hybridization specificity of the designed arrays was further evaluated experimentally using perfect match (PM)/mismatch (MM) probes (Deng, He et al. 2008). A set of 938 PM probes and a corresponding set of 938 MM probes each for the gram negative bacterium, *Desulforvibrio vulgaris* Hildenborough (*DvH*) (GC content ~63%), and the gram positive bacterium, *Clostridium cellulolyticum H10* (*H10*) (GC content ~37%) were added to a modified GeoChip 5.0S. Each MM probe was generated by dividing the PM probe into 5 equal segments, and then one mismatch was randomly introduced into each segment (Deng, He et al. 2008), for a total of 5 mismatches (10% difference) in each MM probe. The hybridization signals from the MM probes should represent non-specific cross-hybridization (i.e. background noise) to their corresponding PM probes (Deng, He et al. 2008). Previous studies suggested that any probes with a signal intensity ratio of PM/MM > 1.3 would be considered a positive hybridization signal (Hazen, Dubinsky et al. 2010). To test specificity, equal amounts (100 ng) of pure culture DNAs were mixed, labeled and hybridized in triplicate with the modified

GeoChip 5.0S. Under the hybridization conditions used (67 ˚C and 10% formamide), the majority of probes (96.8% for *DvH* and 95.1% for *H10*) had PM/MM ratios larger than 10 (**Figure 3.2**). Not a single PM/MM probe had a PM/MM ratio less than 1.3, and a very small portion (0.8% for *DvH*, and 1.2% for *H10*) of the PM/MM probes had PM/MM ratios less than 5. These results suggested that the background noise due to cross-hybridization is very small under the hybridization conditions used, and hence the designed arrays are highly specific.

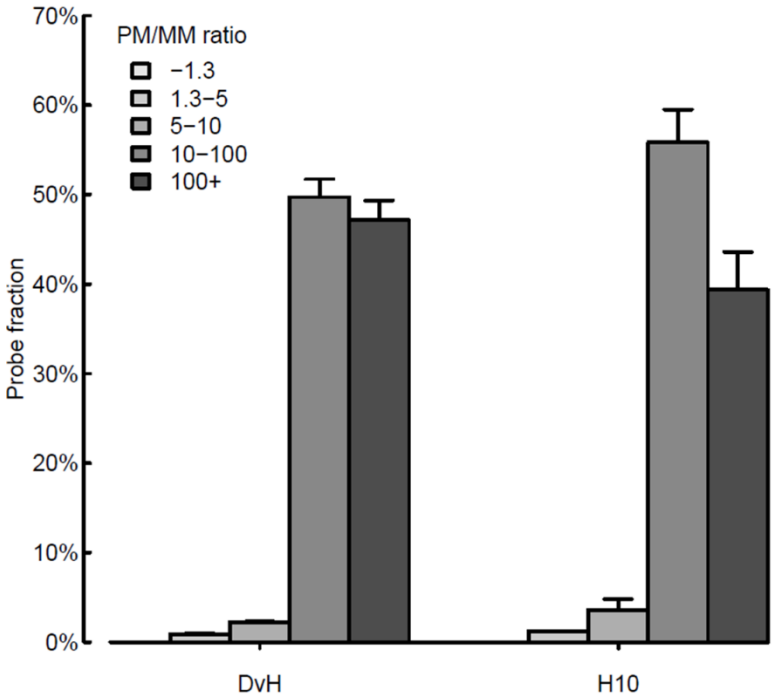


Figure 3.2 Experimental evaluation on the specificity of designed arrays with the perfect match (PM)/mismatch (MM) strategy. 100ng genomic DNAs was labeled with Cy5 and hybridized with a modified GeoChip 5.0S in triplicates. For each PM or MM pair probe, the net signal intensity was obtained by subtracting the signal intensity from Agilent negative spots within a sub-array from the raw signal intensity. The ratio for pair of PM-MM probes was estimated.

3.4.5 Sensitivity of the designed arrays

The hybridization sensitivity of the designed arrays was evaluated with genomic DNAs from *DvH* and *H10*. Pure culture DNAs (0.05, 0.1, 0.5, 1, 5, 10, 50, and 100 ng) were

mixed with soil DNAs from a grassland so that the total amounts of DNAs used for hybridization were all equal to 1,000 ng. The mixed DNAs were directly labeled with Cy5 and hybridized in triplicate with the GeoChip 5.0S containing ~1000 probes from *DvH* and *H10* as described above.

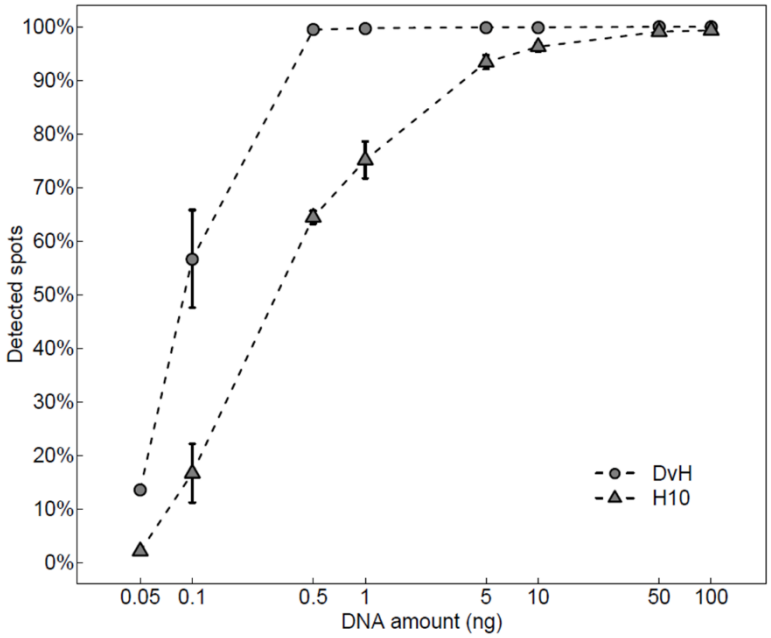


Figure 3.3 Sensitivity evaluation of the designed arrays with pure genomic DNAs. Various amounts of genomic DNAs from *DvH* and *H10* (0.05 ng - 100 ng) were mixed with community DNAs from grassland soils, labeled with Cy5 and hybridized GeoChip 5.0S in triplicate. GeoChip 5.0S contained 938 probes from *DvH* and *H10* respectively.

As shown in **Figure 3.3**, more than 90% (~932) of the pure culture probes were detected at a genomic DNA concentration of 0.5 ng (0.05% of the total community) for *DvH* and 5 ng (0.5%) for *H10*. Over 50% of the probes showed positive hybridization at a genomic DNA concentration of 0.1 ng (0.01%) for *DvH* and 0.5 ng (0.05%) for *H10*. A small percentage of probes (13.6% for *DvH*, 2.1% for *H10*) were still detected for both *DvH* and *H10* at the lowest concentration of 0.05 ng (0.005%). For the low GC content organism (*H10*), hybridization sensitivity is roughly about 10 times lower than the high GC organism (*DvH*). Taken together, these results suggested that the designed

Agilent arrays are highly sensitive, with a detection limit as low as $5\times10^{-4} \sim 5\times10^{-5}$ populations within a complex soil community.

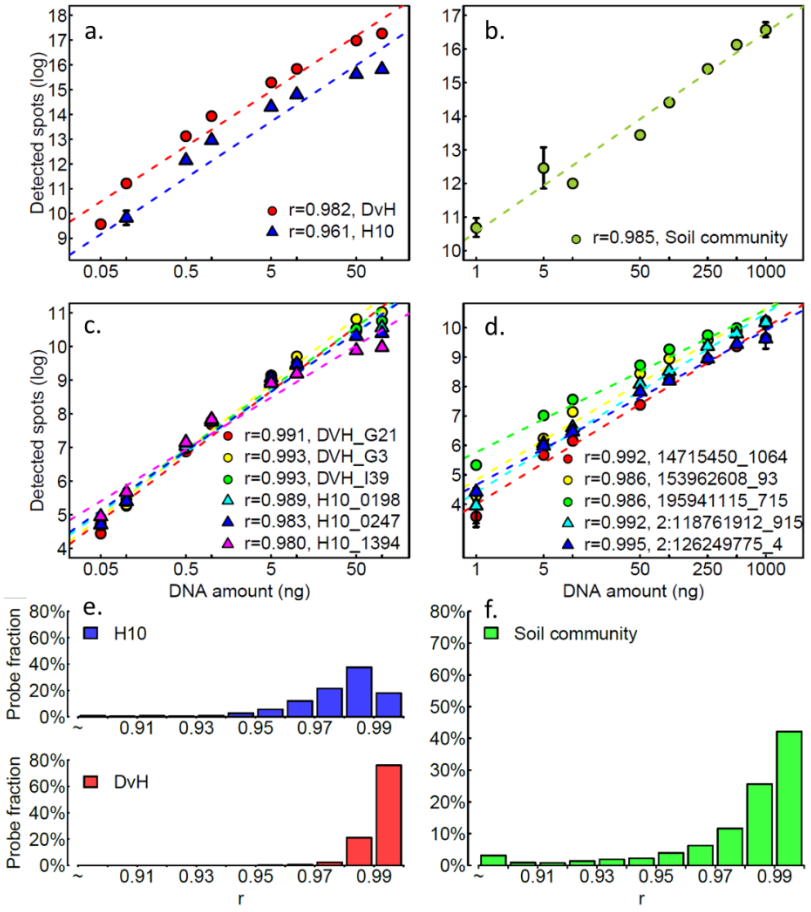


Figure 3.4 Quantitative evaluation of the designed arrays with pure culture and soil community DNAs. Various amounts of pure culture DNAs (0.05, 0.1, 0.5, 1, 5, 10, 50, and 100 ng) and soil community DNAs (1, 5, 10, 50, 100, 250, 500 and 1000 ng) were mixed with different amounts of background DNAs (soil DNAs and Salman sperm DNAs, respectively) so that the total amounts of DNAs are all equal to 1,000 ng. The signal intensity for each spot was corrected by deducting the signal from Agilent negative control, and any spots with 0 or negative values were discarded. A total of 937 and 877 spots for *DvH* and *H10* were included in this analysis respectively. (**a**) Relationship of total signal intensity over all detected spots to the amount of pure culture DNAs used. (**c**) Relationship of total signal intensity for selected representative spots to amount of pure culture DNAs used; (**e**) Distribution of determination coefficients (Pearson correlation coefficient, ρ) based on individual spots for pure culture detection. (**b**) Relationship of total signal intensity over all detected spots to the amount of soil community DNAs used. (**d**) Relationship of total signal intensity for selected representative spots to amount of soil community DNAs used; (**f**) Distribution of determination coefficients (Pearson correlation coefficient, ρ) based on individual spots for soil community detection.

3.4.6 Quantitation of the designed arrays

The quantitative capability of the designed arrays was first evaluated with pure cultures of *DvH* and *H10* in the presence of soil DNAs as background (see sensitivity test). Both signal intensity and DNA concentrations were log transformed. The total signal intensity for all genes was highly correlated with the total DNAs used in hybridization for both *DvH* (Pearson correlation coefficient, $r = 0.982$) and *H10* ($r = 0.961$) (**Figure 3.4**a). Also, all of the individual genes detected showed significant correlations ($r = 0.824$-0.999; *p-value* < 0.05) with DNA concentrations in at least more than three orders of magnitude. Extremely strong correlations between signal intensity and DNA concentrations were observed for some representative genes (**Figure 3.4**c). In addition, 937 *DvH* and 877 *H10* genes were detected in at least 6 concentrations, and about 99% of these detected genes had $r > 0.9$ (**Figure 3.4**e). These results indicated that the hybridization of the designed arrays is highly quantitative with pure culture DNAs in the presence of soil DNAs as background.

The quantitative nature of the designed arrays was also assessed directly with soil DNAs. Different amounts of soil DNAs from a grassland (1, 5, 10, 50, 100, 250, 500 and 1,000 ng) were mixed with Salman sperm DNAs as a background to make up 1,000 ng DNA in total. The mixed DNAs were directly labeled with Cy5, and hybridized with GeoChip 5.0S. As with pure culture DNAs, strong correlations were observed between the total signal intensity and DNA concentrations used for hybridization (**Figure 3.4**b). A total of 2,496 genes were detected in the two highest concentrations (500 ng and 1,000 ng) and across at least 4 of the rest of 6 concentrations. Among these, all of the genes detected on the GeoChip 5.0S showed significant correlations (p-value < 0.05)

between their signal intensity and DNA concentrations across at least three orders of magnitude. Some genes even showed almost perfect correlations (**Figure 3.4**d). About 97% of the genes had $r > 0.9$ (**Figure 3.4**f). Altogether, the above results suggest that the GeoChip hybridization with complex soil DNAs is also highly quantitative across a dynamic range of at least three orders of magnitude.

3.4.7 Application of GeoChip 5.0 to analysis of contaminated groundwater microbial communities

To demonstrate the usefulness of the developed FGAs, we examined the impacts of heavy metal contamination on groundwater microbial communities at the Department of Energy (DOE) Field Research Center (FRC) in Oak Ridge (TN, USA). The groundwater at this field site is heavily contaminated with radionuclides, dissolved organic matter, and nitric acid emitted during nuclear weapon development and processing. A total of 12 wells were selected, representing 4 different groups of contamination levels: no contamination (L0), low contamination (L1), intermediation contamination (L2), and high contamination (L3). A total of 41 physical, chemical, and biological variables were measured, such as heavy metals (e.g. uranium), pH, nitrate, and sulfide (**Table S 2**). Both DCA and clustering analysis showed that different groups of wells were distinctly different among them, and highly similar within individual groups (**Figure S 5**), indicating that the geochemistry and contaminants are quite different among these wells.

Table 3.4 Impact of contamination on the functional gene diversity and evenness. Functional gene diversity for each sample was estimated with Shannon Index, Simpson Index and functional gene richness, and was averaged for each contamination level and compared with each other. A consensus rank of functional gene diversity among three methods was given in the rightest column. Welch's *t*-test for the difference between functional gene diversities and evenness of each pair of contamination levels. Functional gene diversity for each sample was estimated with Shannon Index, Simpson Index and functional gene richness. Statistical significance level was p-value equal to 0.05 or below. The significant testing results were marked in red.

Contamination Level	Richness		Shannon Index		Simpson Index		Evenness	
L0	52495 ± 2631.3		10.85 ± 0.05		0.9999806 ± 9e-07		0.99928 ± 7.1e-05	
L1	44517 ± 959.6		10.69 ± 0.02		0.9999772 ± 5e-07		0.99938 ± 1.1e-05	
L2	40166 ± 3298.3		10.59 ± 0.08		0.9999747 ± 2e-07		0.99942 ± 1.1e-04	
L3	26112 ± 1459.8		10.16 ± 0.05		0.9999612 ± 2e-07		0.99951 ± 3.9e-04	
Sample Grouping	Welch's *t*-test							
	Richness		Shannon Index		Simpson Index		Evenness	
	t	*p*	*t*	*p*	*t*	*p*	*t*	*p*
L0 vs. L1	4.934	0.023	5.277	0.015	5.593	0.011	2.348	0.136
L0 vs. L2	5.061	0.008	4.879	0.013	4.588	0.022	1.891	0.142
L0 vs. L3	15.2	<0.001	16.187	<0.001	14.185	0.001	0.996	0.418
L1 vs. L2	2.19	0.141	2.151	0.149	2.103	0.157	0.692	0.559
L1 vs. L3	18.24	<0.001	15.378	0.001	12.421	0.004	0.575	0.623
L2 vs. L3	6.75	0.008	7.512	0.002	7.813	0.001	0.367	0.744

Since a very low amount of community DNAs were obtained from the highly-contaminated wells, a small amount of community DNAs (5-10 ng) were amplified with Phi 29 (Wu et al. 2006). Then, all of the amplified DNAs (~2 µg) were hybridized to GeoChip 5.0M. A total of 20,295 genes were detected across all samples, varying significantly across different samples. As expected, both functional gene richness and Shannon-Wiener diversity decreased significantly as contamination increased (**Table 3.4**).

Microbial community functional structure was also quite different among these samples as shown in the DCA ordination plots (**Figure S 6**). Samples from each of the group wells, the background, low, moderate and high contaminant wells, were clustered

together but well separated from each other (**Figure S 6**). These results indicated contaminants have great impacts on the functional structure of groundwater microbial communities.

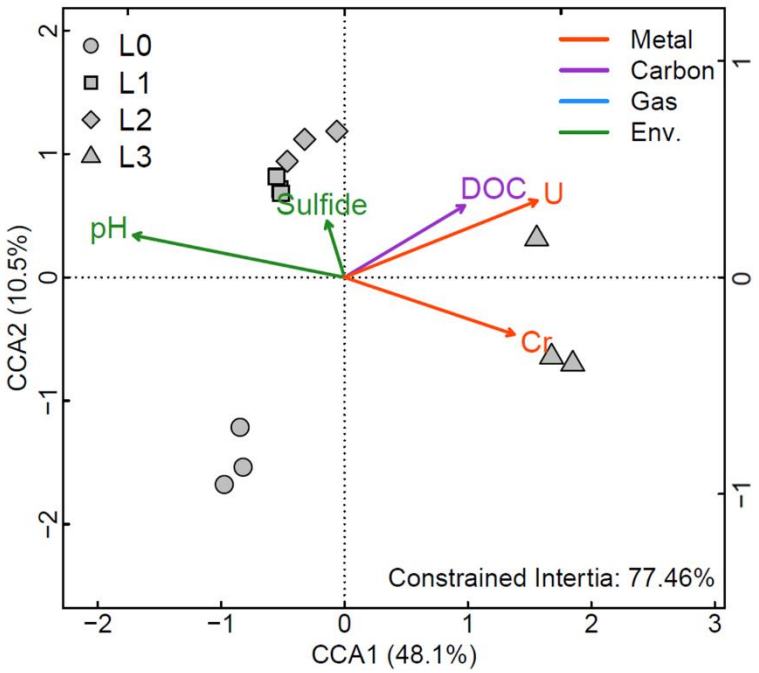
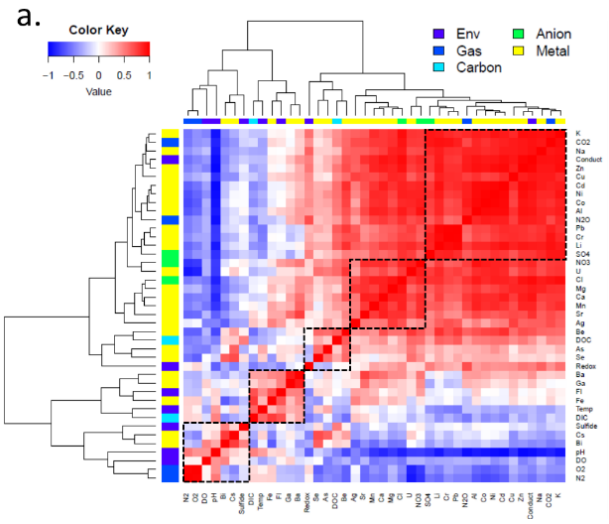


Figure 3.5 CCA on the selected environmental factors and microbial functional gene structure. Top two axis (CCA1 and CCA2) were included, which accounted for 48.1% and 10.5% microbial functional gene structure variation, respectively. A total of 5 environmental factors (U, pH, Cr, Sulfide and DOC) were selected from 41 measured variables based on correlation analysis, and 77.46% CCA inertia was constrained by the selected factors.

CCA analysis was performed to further understand what environmental variables controlled the groundwater microbial community structure (**Figure 3.5**). In this study, a total of 41 environmental variables were measured and subjectively divided into 5 major categories: environmental parameters (env. parameters), gas TCD, dissolved carbon (C), anion and metal ion (ENIGMA web site). Among these, many were highly correlated with each other and five major variable clusters were identified based on the correlation analysis (**Figure 3.6**a). A total of 5 variables, U, pH, Cr, Sulfide and DOC, were selected as representatives for the five clusters (boxed in **Figure 3.6**a) in

subsequent CCA analysis. CCA results showed the differences in the functional gene composition groundwater microbial communities were significantly (*p-value* < 0.001) correlated with changes in the selected variables (**Figure 3.5**). These selected variables could explain up to ~75% of total variations. Partial CCA analysis showed that U and DOC play critical roles in shaping microbial community structure (**Figure 3.6**b).



b.

Variable	Controlled Variables	Constrained Inertia	P-value
U	pH, Cr, Sulfide, DOC	13.77%	0.01
pH	U, Cr, Sulfide, DOC	3.24%	0.613
Cr	U, pH, Sulfide, DOC	6.22%	0.092
Sulfide	U, pH, Cr, DOC	7.48%	0.051
DOC	U, pH, Cr, Sulfide	10.73%	0.022
All 5	none	77.46%	0.001

Figure 3.6 (a) Heatmap of correlation (Pearson correlation) matrix among all environmental factors. The original values of conductivity, Cl, NO,3 SO4, Ag, Al, As, Ba, Be, Bi, Ca, Cd, Co, Cr, Cs, Cu, Fe, Ga, K, Li, Mg, Mn, Na, Ni, Pb, Se, Sr, U and Zn were log transformed due to the nature of the measurements. Factor clusters identified by hierarchical cluster analysis was boxed by the dashed black lines. (b) Partial CCA on the selected environmental factors and microbial functional gene structure. The significant models were marked in red.

3.5 Discussion

Although development and application of high-throughput metagenomics technologies (e.g. next generation sequencing, arrays, mass spectrometry-based proteomics) have revolutionized the capabilities for microbiologists to analyze microbial communities in the environment, various experimental and computational challenges still exist and further advances are needed (Zhou, He et al. 2015). Thus, in this study, we have developed a new generation of functional gene arrays (GeoChip 5.0) which contain

161,961 probes covering functional groups involved in microbial carbon (C), nitrogen (N), sulfur (S), and phosphorus (P) cycling, energy metabolism, antibiotic resistance, metal resistance/reduction, organic contaminant remediation, stress responses, pathogenesis and virulence as well as markers specific for viruses, protists, and fungi. To the best of our knowledge, this is the most comprehensive functional gene arrays currently available for studying microbial communities important to biogeochemistry, ecology, environmental sciences as well as human health.

Compared with previous generations of GeoChip, GeoChip 5.0 has several improved features. First, several new functional categories were included, such as microbial defense, protist, plant growth promotion, pigments and metabolic pathways, to expand our ability to study the associated functional processes. Second, GeoChip 5.0 has a more comprehensive coverage in terms of the number of functional gene families and the number of targeted genes. The vast expansion of functional gene families will allow researchers to analyze more functional processes in more complex ecosystems. In addition, GeoChip 5.0 is *in situ* synthesized by Agilent with smaller spots with higher density (**Table 3.3**). All these distinct features make GeoChip 5.0 a more sensitive and comprehensive tool for analyzing complex microbial communities, and linking their composition with environmental variables and ecosystem functions.

Specificity is one of most critical parameters for detection and is particularly important for analyzing complex environmental samples such as soils because there are numerous homologous sequences for each gene present in a sample. To achieve appropriate specificity, we used specific design criteria to improve specificity. First, seed sequences of each gene family were carefully selected to build HMM model to confirm the targets

for probe design, which should help exclude some of non-target sequences retrieved through bulk automatic downloading at the early stage of GeoChip 5.0 development. Second, multiple experimentally determined criteria based on sequence identity, continuous stretch length and free energy were simultaneously applied to probe selection to ensure that the selected probes have the highest specificity (He, Wu et al. 2005, Li, He et al. 2005). Third, both sequence-specific and group-specific probes were designed using experimentally evaluated and established criteria (He, Wu et al. 2005, Liebich, Schadt et al. 2006). The specificity of each selected probe was verified again by searching against the NCBI databases. Probes from previous GeoChips were verified by searching against the updated GenBank databases. By implementing the above quality control protocols, the final probe sets should be highly specific as demonstrated by the computational evaluation, which showed that only a very small portion (5%) of the designed probes were very close to the criterion thresholds, consistent with previous GeoChip versions (He, Wu et al. 2005, He, Gentry et al. 2007, He, Deng et al. 2010, Tu, Yu et al. 2014). Our experimental evaluation based on PM/MM strategy showed considerable differences (>95% PM probe signal intensities are 10-fold higher) of signal intensity between PM probes and MM probes for both high and low GC content genomic DNAs. Collectively, these results suggest that the probe design strategy used here and for earlier versions of GeoChip is extremely robust and capable of consistently producing highly specific probes regardless of the microarray platforms (He, Wu et al. 2005, He, Gentry et al. 2007, He, Deng et al. 2010, Tu, Yu et al. 2014).

Sensitivity is another important issue for detection. Due to the differences in printing technologies and hybridization protocols, the Agilent-based functional gene arrays

appear to be more sensitive than previous version of GeoChip (Wu, Thompson et al. 2001, Rhee, Liu et al. 2004, Tiquia, Wu et al. 2004, Wu, Liu et al. 2006, Tu, Yu et al. 2014). For the Agilent-based GeoChip, 0.2-1.0 µg community genomic DNA is enough for direct labeling and hybridization, which is not a problem for the majority of environmental samples. Also, our studies further showed a detection limit as low as $5x10^{-4}$ ~ $5x10^{-5}$ populations can be achieved within a complex soil community. Such detection sensitivity is comparable to quantitative PCR. Collectively, our results indicated that the Agilent-based functional gene arrays are extremely sensitive, and should be sensitive enough for analyzing environmental samples from many habitats such as soils, marine sediments, bioreactors and wastewater treatment plants as demonstrated in many previous studies (Zhou 2009, Zhou, He et al. 2015). If the DNA concentration is extremely low (as low as ~ 10 fg, ~2 bacterial cells), a modified method (Wu, Liu et al. 2006) could be used for amplification that assists with GeoChip hybridization. Although more variations could be introduced when extra steps are involved, the experimental results are still meaningful as demonstrated by the application of GeoChip 5.0 to the analysis of contaminated groundwater microbial communities.

Effective meaningful ecological comparisons across different ecosystems require accurate quantitation of taxon and gene abundances. Thus, quantitation is another most important parameter for any detection technology. Since conventional PCR amplification is used in amplicon-based target sequencing, previous studies demonstrated that target gene sequencing is not or is less quantitative in complex communities (Zhou, Wu et al. 2011, Pinto and Raskin 2012, Tremblay, Singh et al.

2015). This is also consistent with the general consensus that traditional PCR amplification is not quantitative (Suzuki and Giovannoni 1996, Qiu, Wu et al. 2001). In theory, since no conventional PCR is involved in shotgun sequencing of whole communities, it is generally believed that shotgun sequencing should be quantitative (Zhou, Wu et al. 2011, Nayfach and Pollard 2016). However, due to high inherent variations of experimental protocols and uncertainty in selecting bioinformatics tools for analysis (Clooney, Fouhy et al. 2016, Kerepesi and Grolmusz 2016, Nayfach and Pollard 2016), it may be impossible to obtain absolute abundance estimations based on shotgun sequencing data alone (Nayfach and Pollard 2016). Different from sequencing-based approaches, absolute abundance of genes can be estimated based on the signal intensity from array hybridization, which reflects the absolute abundance for the amounts of DNAs used for hybridization. This speculation is supported by the results demonstrated in this study. Highly quantitative results were obtained with both complex soil DNAs (r = 0.985) and the pure culture DNAs (r = 0.995) in the presence of soil DNAs as background. These results are also consistent with previous experimental evaluations with both DNAs and RNAs (Rhee, Liu et al. 2004, Tiquia, Wu et al. 2004, Wu, Liu et al. 2006, Brodie, DeSantis et al. 2007, Gao, Yang et al. 2007, He, Deng et al. 2010).

In summary, the developed GeoChip 5.0 contains ~160K probes, covering ~370K sequences in ~1500 gene families. It is the most comprehensive functional gene array available for dissecting the functional structure of complex microbial communities. Both computational and experimental evaluations demonstrated that the developed Agilent-based GeoChip is highly specific, sensitive, and quantitative for characterizing

microbial community functional composition and structure, and should be a powerful tool for linking microbial communities to various ecosystem functional processes. The developed GeoChip is a powerful tool for rapid, high-throughput, sensitive, quantitative and cost-effective analysis of microbial

communities, and can be used a generic tool to address ecological questions important to human health, agriculture, energy, climate change, ecosystem management, and environmental restoration. As previously discussed (Zhou, He et al. 2015), both sequencing-based open format and array-based closed format have different advantages and disadvantages in terms of specificity, sensitivity, quantitation, resolution, reproducibility and novel discovery. Thus, they should ideally be used in a complementary fashion to address complex ecological questions within the context of ecological, environmental and medical applications (Zhou, He et al. 2015). The functional gene arrays developed here is an important part of the integrated omics toolbox for microbial community analyses.

Chapter 4: The EcoFun-MAP: An Ecological Function Oriented Metagenomic Analysis Pipeline

4.1 Abstract

Functional analysis of deep shotgun metagenomics sequencing is computationally challenging. Here we present an Ecological Function oriented Metagenomic Analysis Pipeline (EcoFun-MAP), to facilitate analysis of shotgun metagenomic sequencing data in microbial ecology studies. The EcoFun-MAP consists of reference databases of different data structures, with a selective coverage of functional genes that are important to ecological functions. Meanwhile, multiple predefined data analysis workflows were built on the databases with most updated bioinformatics tools. Furthermore, the EcoFun-MAP was implemented and deployed on High-Performance Computing (HPC) infrastructure with high accessible and easy-to-use interfaces. In our evaluation, the EcoFun-MAP was found to be fast (multi-million reads/min.) and highly scalable, and capable of addressing disparate needs for accuracy and precision. In addition, we showcase the effectiveness of the EcoFun-MAP by applying it to reveal differences among metagenomes from underground water samples, and provide insights to link the metagenomic differences with distinctive levels of contaminants. The EcoFun-MAP is open for public use and can be found available at our website: http://zhoulab5.rccc.ou.edu:7999.

4.2 Introduction

Next generation sequencing (NGS) technology has revolutionized metagenomics and microbial ecology studies due to immense improvements made in sequencing speed, throughput, and cost. It can produce a formidable number of raw reads during a single run, which allows in-depth profiling of microbial community from an environmental sample and leads to novel discoveries of microbial species. As NGS technology has been democratized to microbial ecologists for the past decade, numerous metagenomics studies have been enabled to investigate microbial composition, diversity, function, dynamics and interaction in diverse and complex environments. Recent remarkable examples include studies of microbial communities from tundra (Xue, M. Yuan et al. 2016) and desert (Rasuk, Fernández et al. 2016) soil, Arctic marine sediments (Algora, Vasileiadis et al. 2015), deep-sea hydrothermal vents (Topçuoğlu, Stewart et al. 2016) and various hosts (e.g. coral reef). Indeed, NGS technology has facilitated microbial ecology studies in an unprecedented way, improved our understanding and mechanistic modeling of microbial community (Franzosa, Hsu et al. 2015), and established links between diversity and composition of microbial community and the biogeochemical state of ecosystem (Schimel 2016).

NGS technology for metagenomics has two major applications: amplicon sequencing and shotgun metagenomics sequencing (Scholz, Lo et al. 2012). Amplicon sequencing typically relies on PCR amplification and can target both 16S rRNA genes and specific functional genes. It can quickly obtain a distribution profile of taxonomy or certain functional genes of an environmental sample at acceptable cost, but it has several limitations of which researchers have become more and more aware (Hong, Bunge et al.

2009, Sharpton, Riesenfeld et al. 2011, Wylie, Truty et al. 2012, Langille, Zaneveld et al. 2013, Logares, Sunagawa et al. 2014, Zhou, He et al. 2015). Most of those limitations stem from the selective PCR amplification, which may lead to partial, skewed or inaccurate profiles of microbial communities. Instead of being subject to limitations, shotgun metagenomics sequencing avoids selective PCR, thus it is capable of recovering most parts of a metagenome from an environmental sample, including both taxonomically and functionally informative fragments. Therefore, shotgun metagenomics sequencing may address the question of not only who are there, but also what they can do. This is highly desirable, because it paves a way to meet one of the most important goals in the field of microbial ecology: understanding and modeling microbial community despite the extreme diversity and complexity of the community, which is hardly possible to achieve without addressing both questions at the same time. However, fully revealing microbial community functional composition in samples from complex systems, such as soil, may require a commensurate depth of shotgun metagenomics sequencing, which brings computational challenges, especially to microbial ecologists. First, the sheer amount of shotgun metagenomics sequencing data is difficult to handle. A single sample can have the data size that is not convenient to do normal file operations, and metagenomics projects nowadays can have hundreds of samples, which may generate overwhelmingly big volume of data and easily cause trouble to store, manage and share. Second, processing and analysis of shotgun metagenomics sequencing data are complex and computationally intensive. A typical workflow may have major steps including quality assessment, poor read or base trimming, assembly (optional), binning (optional), gene prediction (optional), and

annotation. Each step is complicated by tool selections and detailed tool settings, and computational complexity of the step can be further dependent on multiple factors, such as tool algorithm, reference database size, sequencing platform (Oulas, Pavloudi et al. 2015) and effectiveness of previous steps. Since delivery of increasingly large data volume by future NGS technology is expected, advances in informatics and computational resources are crucial to meet the requirement of ease and efficiency of functional metagenomics analysis, without which shotgun metagenomics studies will exhaust available resources and be greatly impeded by the computational bottlenecks. Propitiously, development of computational tools and database is rapid. Those frequently used computational resources for shotgun metagenomics analysis are available and provided in different forms, including standalone programs for specific steps of analysis (e.g. FastQC (Andrews 2010), Btrim (Kong 2011), LUCY2 (Li and Chou 2004), ABySS (Simpson, Wong et al. 2009), Meta-IDBA (Peng, Leung et al. 2011), MetaVelvet (Namiki, Hachiya et al. 2012), IDBA-UD (Peng, Leung et al. 2012), Prodigal (Hyatt, Chen et al. 2010), FragGeneScan (Rho, Tang et al. 2010), NCBI BLAST (Altschul, Madden et al. 1997), BLAT (Kent 2002), Bowtie (Langmead, Trapnell et al. 2009, Langmead and Salzberg 2012) and Diamond (Buchfink, Xie et al. 2015)), reference databases (e.g. NCBI NT and NR (Pruitt, Tatusova et al. 2005, Clark, Karsch-Mizrachi et al. 2016), KEGG (Kanehisa, Araki et al. 2008), eggNOG (Powell, Forslund et al. 2013), PFAM (Finn, Bateman et al. 2013) and SEED (Overbeek, Olson et al. 2014)), integrated analysis pipelines (e.g. IMG/MER (Markowitz, Chen et al. 2014), MG-RAST (Glass, Wilkening et al. 2010), CAMERA (Seshadri, Kravitz et al. 2007) and Parallel-META (Su, Pan et al. 2014)). While these resources are becoming

more powerful and efficient, there are still barriers hindering shotgun metagenomics analysis in microbial ecology. First, lack of computational skills and access to advanced computing hardware may cause difficulties for microbial ecologists taking advantage of standalone programs and databases. Even installation and configuration of these tools and databases can result in a non-trivial amount of work for data analysis novice. Second, reference databases are usually for general annotation purpose. As being general, those databases are inclusive to annotation needs from distinctive disciplines, but it can cause unnecessary computing cost, especially when the sizes of the databases is becoming exponentially large due to the explosion of sequencing project submission facilitated by NGS technology. Third, integrated analysis pipelines, particularly web-based pipelines, are more accessible to users with less computational background, but most of the available pipelines are only offering graphic user interfaces or automatic solutions building upon aforementioned standalone tools and reference databases, which may be under-optimized or lack focus and efficiency for functional metagenomics analysis in field of microbial ecology. Nevertheless, few tools provided efficient and accessible solutions for functional analysis of shotgun metagenomics data with a clear focus on linking functional composition of microbial community to ecological functions and geochemical processes.

Here to ease functional analyses of shotgun metagenomics sequencing data derived from typical microbial ecology studies, we developed an Ecological Function-oriented Metagenomics Analysis Pipeline (EcoFun-MAP), which is designed based on a functional gene-centric paradigm. To develop the EcoFun-MAP, we first carefully defined the coverage of the EcoFun-MAP by only including functional categories and

genes that were important to ecological functions and geochemical processes. Based on the coverage, then we collected and curated relevant nucleotide and amino acid sequences, and constructed reference databases of different data structures. Next, several data processing workflows were designed and build on the databases and different tools to provide flexible analysis for addressing disparate needs for speed or sensitivity. Then, the EcoFun-MAP was implemented on the basis of High-Performance Computing (HPC) infrastructure with web-based user interfaces. In this study, we also evaluated speed, accuracy, and precision of the EcoFun-MAP, and demonstrated its effectiveness by applying it to analyze metagenomes from underground water samples from wells where different levels of contaminants are present.

4.3 Material and methods

4.3.1 Selection of functional categories and genes

We defined applicable scope of EcoFun-MAP and organized it into 15 major categories (detailed description and selection rationale in supplementary text) associated with geochemical processes and ecological functions that are important to environmental metagenomics studies, including Carbon (C), Nitrogen (N), Sulfur (S), and Phosphorus (P) cycling, antibiotic resistance, organic contaminant degradation, metal homeostasis, stress response, microbial defense, electron transferring, plant growth promotion, virulence, protist, virus and others (metabolic pathways, pigment biosynthesis and *gyrB*). We then selected and further categorized functional genes based on their roles in the major categories. Finally, three to four levels of organization have been generated for selected functional genes. The highest level is the most general class, which is one of the 15 major categories, the lowest level is the most specific class, which is the

functional genes themselves, and in between are the primary subcategory and secondary subcategory. For example (**Figure S 7**), the C cycling category (144 genes) consists of three primary subcategories, including C degradation (60 genes), C fixation (61 genes) and Methane (23 genes). The primary subcategory of C degradation has 18 secondary subcategories (e.g. Starch degradation, Cellulose degradation and Lignin degradation), the C fixation has 8 secondary subcategories (e.g. Calvin cycle, Dicarboxylate/4-hydroxybutyrate cycle and 3-hydroxypropionate bicycle), and the Methane has two secondary subcategories (i.e. Methane oxidation and Methanogenesis). Each secondary subcategory has a number of genes ranged from 1 to 21 (**Figure S 7**).

4.3.2 Retrieval of functional gene sequences

National Center for Biotechnology Information (NCBI) Entrez databases (Coordinators 2013) were used as the source to retrieve functional gene sequences for constructing EcoFun-MAP databases. We manually crafted keyword-based query for each functional gene, and submitted it programmatically to the Entrez databases to search and retrieve both protein and nucleotide candidate sequences via Entrez Programming Utilities (E-utilities) (Coordinators 2013). A typical such a query has been designed to consist of all aliases and variants names of the corresponding gene known to us, as well as other NCBI search constraints (e.g. organism), braces and logic operators (e.g. AND, OR and NOT). By carefully crafting the keyword-based query, relevance of research results can be improved as number of the results drop, therefore initial quality control can be achieved before the EcoFun-MAP database construction and computational cost can also be reduced for later processing. For example, a keyword based query for *nifH* gene (**Figure S 8**) has returned 34,077 nucleotide records and 31,522 protein records, which

were much less than 100,728 nucleotide records and 82,722 protein records in total returned by simply using "nifH" as the search query (retrieval test date: Jan. 23rd, 2017), and successfully excluded irrelevant records, such as Sinorhizobium sp. partial *nodA* gene (GenBank ID: Z95242.1) and Heliobacterium gestii partial *anfH* gene (GenBank ID: AB100834.1). Next, from records retrieved using keyword based query search, a minimum of 5 to a few hundred random seed sequences were selected manually on the basis of two criteria: 1) seed sequences must be experimentally confirmed in literature, and 2) seed sequences must be distinctive from each other. Finally, redundant records (i.e. records with identical GenBank ID and description) were removed. To this end, candidate sequences and seed sequences have been prepared for each selected EcoFun-MAP gene and ready for EcoFun-MAP database construction.

4.3.3 Construction of EcoFun-MAP databases

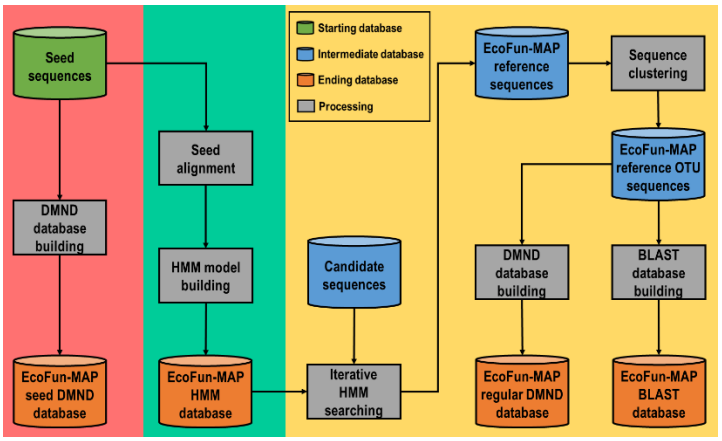


Figure 4.1 The flowchart of construction of databases/datasets in development of the EcoFun-MAP. Cylinders represent starting (green), intermediate (blue) and ending (orange) databases. Grey rectangles represent processing steps in construction, which take content of databases or output of immediate upstream processing steps as input for processing.

EcoFun-MAP databases were constructed using aforementioned candidate sequences and seed sequences. The construction workflow (**Figure 4.1**) has produced four ending

databases, including a seed sequence based DIAMOND index database (EFM-DI-DB-S), a Hidden Markov Model (HMMs) based database (EFM-HMM-DB), a functional gene reference sequence based DIAMOND index database (EFM-DI-DB-R) and a functional gene reference sequence based NCBI-BLAST index database (EFM-BLAST-DB). These four databases are differed in size and data structure, thus allow different processing speed and accuracy. To build the databases, first, the protein seed sequences for all covered functional gene families were pooled together and used for directly building the EFM-DI-DB-S. Meanwhile, the seed sequences of each functional gene family were aligned and the resulting alignment was used for building the EFM-HMM-DB. Next, reference sequences of each functional gene family were selected from the protein candidate sequences by iteratively searching the candidate sequences against the EFM-HMM-DB. The iterative searching has following steps: 1) set up an initial e-value cutoff, 2) searching the candidate sequences against the EFM-HMM-DB using the e-value cutoff, and 3) manually evaluate the resulting candidate sequences passed the searching, and adjust e-value cutoff for repeating the searching if needed. Due to the different set of sequences that each gene has, the best cutoff value for selecting reference sequences could differ among genes, therefore manual effort to make repeated adjustments is vital to ensure the quality of reference sequences. Finally, we clustered the reference sequences of each functional gene family into multiple Functional Clusters (fClusters) based on the sequence similarity, which were used for building both EFM-DI-DB-R and EFM-BLAST-DB.

4.3.4 Design of EcoFun-MAP workflows

By taking advantage of disparate databases that have been constructed, a total of 5 EcoFun-MAP workflows have been designed, which were labeled as ultra-fast, fast, moderate, sensitive and ultra-sensitive mode (**Figure 4.2**), respectively. All the workflows used the same procedure for preprocessing raw sequencing reads, in which quality trimming and gene prediction took place one after another. The preprocessing procedure should remove bases of low quality or ambiguity and reads of overly short length, and identify and extract gene fragments from input reads. Then route for further analyzing the preprocessed reads diverged to form the 5 modes. In the ultra-fast mode, the preprocessed reads were directly searched against the EFM-DI-DB-S database. The fast mode workflow extended the ultra-fast mode by further searching the EFM-DI-DB-S annotated reads against the EFM-HMM-DB and then searching the resulting reads against the EFM-BLAST-DB. Similarly, in the moderate mode, the preprocessed reads were directly searched against the EFM-DI-DB-R. The sensitive mode workflow extended the moderate mode by further searching the EFM-DI-DB-R annotated reads against the EFM-HMM-DB and then searching the resulting reads against the EFM-BLAST-DB. Finally, in the ultra-sensitive mode workflow, the preprocessed reads were first searched against the EFM-HMM-DB, and then searching the resulting reads against the EFM-BLAST-DB. In the end, all workflows provided an optional step to normalize counts of hits based on the average length of reference sequences from the gene families of the hits. The designed workflows relied on different databases and processing steps and should provide distinctive performance in terms of both speed,

accuracy, and precision, therefore allowed needed flexibility for data analysis in practice.

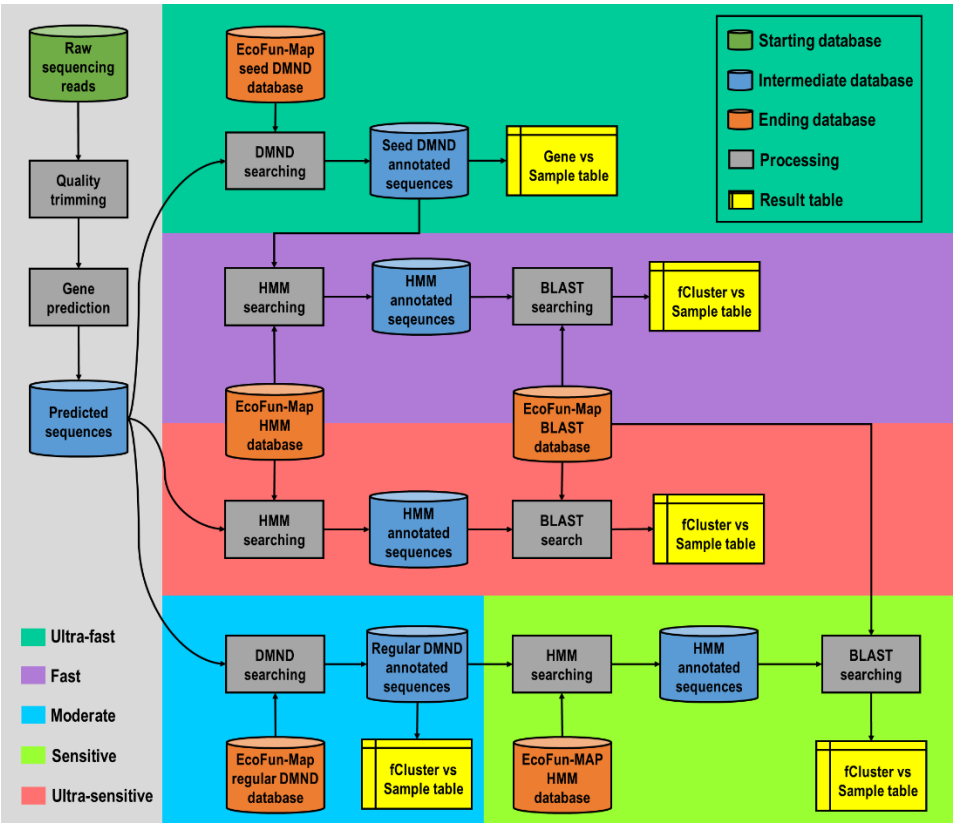


Figure 4.2 The flowchart of five workflows in the EcoFun-MAP, which include mode of Ultra-fast (green background), Fast (purple background), Moderate (cyan background), Sensitive (green background), and Ultra-sensitive (red background). The preprocessing steps are on the grey ground. Cylinders represent starting (green), intermediate (blue) and ending (orange) databases. Grey rectangles represent processing steps in construction, which take content of databases or output of immediate upstream processing steps as input for processing. Shapes of yellow documents represent resulting matrix-like table.

4.3.5 Experimental datasets

Experimental datasets for showcasing and evaluating the EcoFun-MAP were sequenced from underground water samples from the Oak Ridge Integrated Field Research Challenge site (OR-IFRC; Oak Ridge, TN; http://www.esd.ornl.gov/orifrc/) (Chiachi Hwang et. al. 2009). The OR-IFRC site was established by the US Department of Energy for researching the long-term treatment of radionuclide wastes, which provides

an ecosystem of extremity for studying microbiomes under gradients of salinity, pH and contaminants including Uranium, nitrate, sulfide, and other heavy metals (Chiachi Hwang et. al. 2009). In this study, we took a total of 12 water samples from underground wells that can be categorized into 4 contamination levels: no contamination (L0), low contamination (L1), intermediation contamination (L2), and high contamination (L3). Three samples were taken for each level. Each sample was processed and microbial community DNA was extracted with the protocol described in previous studies. Metagenome of each sample represented by the extracted DNA was sequenced using the shotgun method with Illumina HiSeq 2000 sequencer. Upon the completion of HiSeq running, about 1,816.7 million of 150 bp raw reads were generated in total, which counted for about 272.5 Gbp data. Data size for each sample is ranged from about 79.6 Gbp (GW199) to about 266 Gbp (FW300). More information about HiSeq output for each sample can be found in supplementary table S2.

4.4 Results

4.4.1 Implementation and deployment of EcoFun-MAP

A number of bioinformatics tools have been used for constructing the EcoFun-MAP databases, as well as developing the workflows. For constructing the databases, the key processing steps including the seed sequence alignment, HMM building, HMM searching, sequence clustering, DIAMOND index building and BLAST index building were implemented using ClustalW (Li 2003), hmmbuild (HMMER3 (Finn, Clements et al. 2011)), hmmsearch (HMMER3 (Finn, Clements et al. 2011)), CD-HIT (Li and Godzik 2006), DIAMOND (Buchfink, Xie et al. 2015) and MAKEBLASTDB (Altschul, Gish et al. 1990), respectively. All of the tools involved in the database

construction were used with default parameters, except the CD-HIT used for sequencing clustering, whose parameter regarding the within-cluster similarity of sequence was set to 95% explicitly.

For developing the workflows, the key processing steps including quality trimming, gene predicting, HMM searching, DIAMOND index searching and BLAST index searching were implemented using Btrim, FragGeneScan+ (Rho, Tang et al. 2010), hmmsearch (HMMER3 (Finn, Clements et al. 2011)), DIAMOND (Buchfink, Xie et al. 2015) and BLASTN (Altschul, Gish et al. 1990), respectively. The workflows have preset parameters for each processing step, and can also accept users' changes on the parameters for meeting specific speed or accuracy needs. For example, the Btrim used in the quality trimming for all the workflows has two major parameters: moving window size and average quality cutoff within the window. The default moving window size was set to 5 and the default average quality cutoff was set to 20 by the EcoFun-MAP, but users can lower the moving window size or set higher the average quality cutoff to increase the quality of trimmed reads.

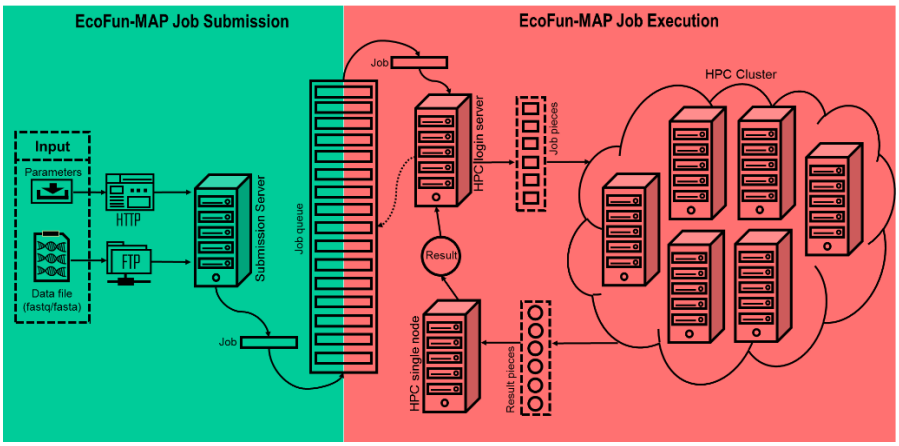


Figure 4.3 The scheme of implementation and deployment of the EcoFun-MAP. Submissions of the EcoFun-MAP jobs (green background) are handled by a standalone server. Further processing and execution of the jobs are performed on a HPC cluster.

The databases and workflows of EcoFun-MAP were deployed on an HPC cluster with a web-based Graphic User Interface (GUI) for access and job submission (**Figure 4.3**). A single EcoFun-MAP job submission requires at the beginning a data file and all parameters that will be used for the selected workflow. EcoFun-MAP provides an FTP application for data file transferring and an HTTP application (website) to accept parameter settings. After being submitted, a job will be scheduled in a job queue and sent to the HPC cluster in a "first in, first out" (FIFO) order for further EcoFun-MAP processing. When executing a job, the HPC cluster will 1) break down the job into small pieces, 2) map job pieces to available nodes, 3) run the selected workflow for the pieces in parallel, and 4) collect and reduce outputs of all pieces, and prepare final result for downloading by the job submitter. The implementation of EcoFun-MAP depends on both open source software and in-house scripts. The FTP application was provided on the basis of installation and configuration of vsftpd (version 3.0.3), the parameter submission website was built using Django, and the job queue was developed using Celery with Redis as the message broker. In-house Perl, Python, Shell and SLURM job scheduling scripts were also used throughout the EcoFun-MAP implementation. Their major functions or roles included the following: 1) job management, 2) calling or executing bioinformatics tools, 3) data file format conversion (e.g. convert FASTQ formatted file into a FASTA one), 4) breaking down, mapping and reducing dataset and 5) data I/O and transferring. At last, the HPC cluster hosting the EcoFun-MAP currently has two types (type I and type II) of computing nodes and each type has 5 nodes, which consists of a total of 10 nodes for handling EcoFun-MAP tasks. The type I node has 24 cores and 64GB RAM each, and the type II node has 24 cores and 128GB RAM each.

The HPC cluster also provides 128TB hard disk space for temporal storage of input, intermediate data and result from tasks of the EcoFun-MAP.

4.4.2 Coverage of EcoFun-MAP

Table 4.1 Overall summary of coverage of the EcoFun-MAP databases by major categories.

Major categories	No. of						
	primary sub-categories	genes	seed sequences	HMM models	fClusters	reference sequences	covered taxonomical IDs
C Cycling	4	138	1,410	209	42,000	156,769	9,462
N Cycling	9	25	639	51	23,530	116,201	7,837
S Cycling	9	26	710	40	9,295	23,123	3,390
P Cycling	4	7	129	12	5,212	18,442	4,620
Organic Contaminant Degradation	9	149	1,290	250	19,139	81,479	7,755
Antibiotic resistance	3	19	234	48	24,450	105,376	6,708
Stress	24	100	2,529	159	47,809	223,627	9,379
Metal Homeostasis	25	120	658	161	5,529	247,826	7,217
Microbial Defense	2	65	495	66	10,613	33,499	3,204
Metabolic Pathways	2	4	54	8	2,110	8,338	2,210
Plant Growth Promotion	3	21	916	21	2,597	17,426	6,344
Pigments	6	26	128	29	1,755	4,398	1,024
Electron transfer	1	12	303	12	1,326	5,468	291
Virulence	42	608	3,467	613	10,415	63,605	4,463
Virus	3	113	984	113	8,079	63,324	15,697
GyrB	1	1	297	13	6,624	37,241	37,241
Protist	13	57	257	57	2,718	11,221	3,076
Total	160	1,491	14,500	1,862	280,247	1,217,363	49,018

The EcoFun-MAP covered 17 major categories and a total of 150 primary subcategories (**Table 4.1**), should be able to provide a comprehensive survey of functional genes important to biogeochemistry, ecology, environmental science, agriculture and public health. The EcoFun-MAP had 1,491 functional gene families, for which 14,500 seed sequences in total were selected and 1,862 HMM models were built. Meanwhile, a total of 1,217,363 reference functional gene sequences were retrieved and confirmed using the iterative HMM searching, which are originated from about 50,000 taxonomical units

that were distinguishable based on their taxonomical IDs. Based on these sequences, more than 280,000 fClusters were generated and further incorporated in the EcoFun-MAP.

4.4.3 Evaluation of speed and accuracy

Speed. Sample FW300, the one with the largest raw data size, was selected from underground water samples for evaluating the speed of the EcoFun-MAP. We randomly drew 5 subsamples in different number of reads (0.7M, 3.5M, 7M, 35M, 70M) from the FW300, which are counted for in different size of raw sequencing data (~100Mbp, ~500Mbp, ~1Gbp, ~5Gbp and ~10Gbp), to evaluate the speed of EcoFun-MAP workflows. Each workflow was run on the subsamples with same hardware configuration (10 nodes and 4 cores on each) and same pipeline parameters. According to the size of each individual reference database and speed of each individual bioinformatics tool that were used in each workflow, we expected that ranks of speed of workflow should be the following: the ultra-fast mode > moderate mode > fast mode > sensitive mode > ultra-sensitive mode.

Table 4.2 Summary of results for evaluating speed of five workflows in the EcoFun-MAP. Subsamples with different number reads randomly drawn from the largest sample, the FW300, are used for the evaluation. Preparing time here refers to the time consumed outside the workflows, including file decompression, data transferring and partitioning and job scheduling.

Mode	Preparing/main processing/total time (s)					Lowest/highest/ average speed (No. of reads in M/min.)
	0.7M reads (~100Mbp)	3.5M reads (~500Mbp)	7M reads (~1 Gbp)	35M reads (~5 Gbp)	70M reads (~10 Gbp)	
ultra-fast	125/60/185	183/121/304	185/180/365	549/540/1,089	978/1,027/2,005	~0.7/4.1/2.5
fast	124/180/304	241/241/482	188/360/548	608/841/1,449	915/1,506/2,421	~0.2/2.8/1.5
moderate	159/60/219	186/180/366	190/240/430	615/602/1,217	862/1,145/2,007	~0.7/3.7/2.1
sensitive	123/242/365	188/300/488	288/361/649	584/1,021/1,605	981/1,865/2,846	~0.2/2.2/1.2
ultra-sensitive	125/181/306	187/485/672	189/840/1,029	617/3,966/4,583	920/7,341/8,261	~0.2/0.6/0.4

In general, the result (**Table 4.2**) showed that our expectation was met, as the ultra-fast mode workflow had the fastest speed, which was finished running on the largest subsample (70M reads) in 1,027 seconds (s), and then was followed by fast (1,506 s), moderate (1,145 s), sensitive (1,865 s) and ultra-sensitive (7,341 s) mode in order of decreasing speed. The running of workflows on the largest subsample yielded the highest speed for all workflows (~0.6-4.1M reads/min.), and speed of workflows increased as the data size went up. The running on the smallest subsample yielded the lowest speed for all workflows (~0.2-0.7M reads/min.). In comparisons of individual workflows, the ultra-fast mode is more than 7 times faster than the ultra-sensitive mode for the largest subsample, but only 3 times faster for the smallest subsample in our test. The result for speed evaluation suggested that the EcoFun-MAP is fast (average speed from ~0.4 to ~2.5 M reads/min.) and highly scalable in high-throughput sequencing data analysis, in which time cost is expected to increase less than linearly as data size hikes, because of the increases in speed.

Accuracy and precision. We also evaluated the EcoFun-MAP in terms of accuracy and precision. It is a non-trivial task to define accuracy in analysis of workflows at the first place, because in general, the true identity of each read is hardly accessible due to possible sequencing error rate, ambiguity hits and change of read alignment score thresholds. In our evaluation, we used the result from the ultra-sensitive workflow as a reference for the comparisons among all workflows, because the ultra-sensitive workflow 1) performs homolog based search for every read, which takes into account information about protein domain structure and thus is considered to be more accurate than read mapping based only on sequence identity, and 2) utilizes probabilistic models

built on multiple sequence alignments and is thus generally more capable of detecting remote homologs than relying on single read. Therefore, the accuracy rate of target workflow was defined as the ratio of number of reads annotated by both the target workflow and the ultra-sensitive workflow to number of reads annotated by the ultra-sensitive workflow. Similarly, the precision rate was then defined as the ratio of number of reads annotated by both the target workflow and the ultra-sensitive workflow to the total number of reads annotated by the target workflow. We further defined four levels of accuracy and precision based on how reads were annotated by both the target workflow and the ultra-sensitive workflow. Level 1, 2, 3 and 4 are used for situations that reads were annotated by both the target workflow and the ultra-sensitive workflow with the same gene, secondary subcategory, primary subcategory, and category, respectively.

Table 4.3 Overall summary of results for evaluating accuracy and precision of five workflows in the EcoFun-MAP. The results here are based on counts of hits from the running of five workflows on all samples.

Mode	Accuracy rate				Precision rate			
	Level 1	Level 2	Level 3	Level 4	Level 1	Level 2	Level 3	Level 4
ultra-fast	70.2%	73.8%	74.5%	77.6%	8.1%	8.5%	8.6%	8.9%
fast	85.4%	88.3%	88.8%	91.9%	2.8%	2.9%	3.0%	3.1%
moderate	69.3%	69.5%	69.5%	69.8%	87.0%	87.2%	87.2%	87.5%
sensitive	84.7%	84.9%	84.9%	85.2%	85.9%	86.0%	86.0%	86.3%

We ran all EcoFun-MAP workflows on the data from aforementioned 12 underground water samples and compared their annotation results for evaluating the accuracy and precision for each workflow. The running of all workflows on about 1816.7 million reads (~272.5 Gbp) in total were completed in less than a week. The result (**Table S 4**) showed that the number of hits produced by workflows was ranged from ~2.1 million (0.12%; moderate mode) to ~81.1 million (4.46%; Fast mode). In general, the fast mode

produced the most hits of all (3.35% - 6.58%) across all samples, the moderate mode produced the least (0.06% ~ 0.27%), and the sensitive mode had very similar yield (0.07% - 0.34%) as the ultra-sensitive mode (**Table S 4**). Upon the definitions of accuracy rate, the result (**Table 4.3**) showed that accuracy rates of other workflows were good in general (~70% above). The fast workflow had the highest accuracy rate of all levels (85.4%, 88.3%, 88.8% and 91.9%) except for the ultra-sensitive workflow, which was then followed by the Sensitive and Ultra-fast mode, and the moderate mode had the lowest (69.3%, 69.5%, 69.5% and 69.8%). Differences of accuracy rate among distinctive levels of accuracy were small (< 0.5%) in the moderate and sensitive mode, and higher in the ultra-fast (~7.4%) and fast mode (~6.5%). Apart from results of accuracy evaluation, the moderate mode had the highest precision rate of all levels (87.0%, 87.2%, 87.2% and 87.5%), which was then followed by the sensitive and ultra-fast mode, and the Fast mode had the lowest (69.3%, 69.5%, 69.5% and 69.8%). Differences of precision rate among distinctive levels of accuracy were small (< 0.5%) in all the modes. The sensitive mode achieved relatively high in both accuracy (~85%) and precision (~86%) rate. The results suggested that performance of the EcoFun-MAP in terms of accuracy and precision depended on the selection of workflow. Detailed results of evaluation of accuracy and precision were provided based on each sample, which showed a similar trend and can be found in **Table S 4**.

4.4.4 Real study application

We analyzed metagenomes from 12 underground water samples based on the running results of the EcoFun-MAP workflows to demonstrate the usefulness of the EcoFun-MAP. Meanwhile, analyses based on SEED Subsystem annotation workflow was

provided for purpose of contrast. Microbial community functional gene composition was compared among the samples as shown in the DCA ordination plots (**Figure 4.4**). The ordination results didn't show drastic differences among all workflows. Samples from group L3 were observed to separate from other groups in all workflows with relatively high within-group distances. Clear separation of L2 samples from other groups were found in the moderate, sensitive and ultra-sensitive workflows. Clear separation of all four groups from each other was only observed in results based on the ultra-sensitive workflow (**Figure 4.4**).

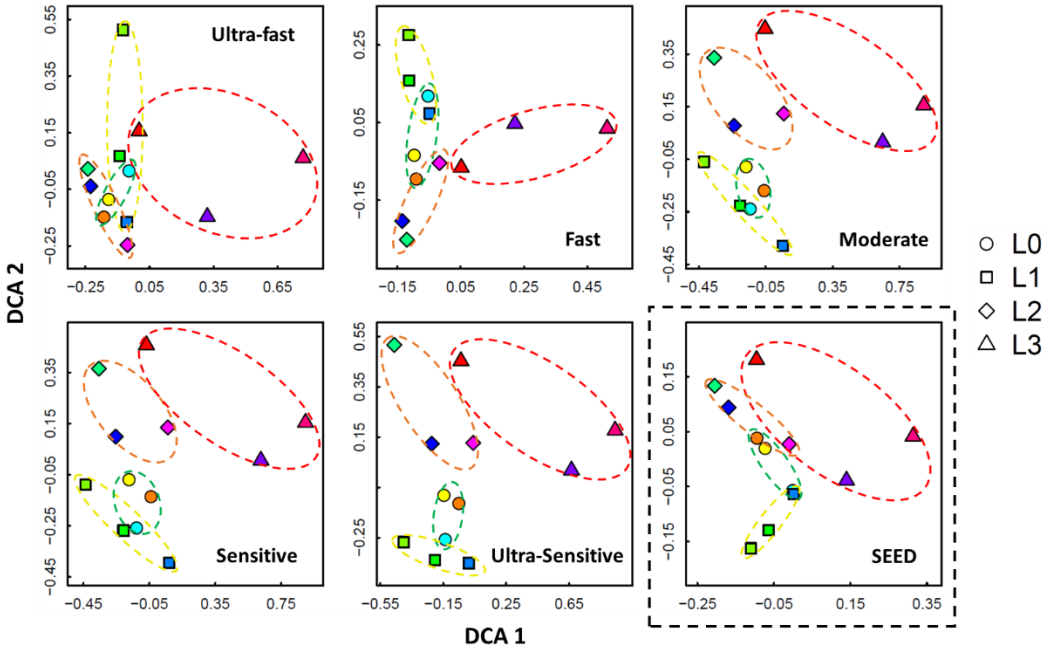


Figure 4.4 Detrended Correspondence Analysis (DCA) of functional gene composition of metagenomes from 12 underground water samples. Analyses of functional gene composition based on results from five workflows of the EcoFun-MAP are provided. Analysis based on result from annotation based on SEED subsystem (boxed by dashed line) is also provided for purpose of contrasting. Each sample is represented by a distinctive color. Cycles, squares, diamonds and triangles are used for showing samples from group of L0, L1, L2 and L3, which are also cycled with green, yellow, orange and red eclipses, respectively.

The analyses of function gene richness in 4 sample groups had similar results for all workflows (**Figure 4.5**). The richness of functional genes was significantly lower (*p*

value < 0.05) in L3 samples than in L0 samples, which was shown in analyses based on all the EcoFun-MAP workflows and SEED Subsystem annotation. The analyses based on the fast mode and SEED Subsystem annotation also showed a significantly lower (*p value* < 0.05) richness of function genes in L3 samples than in L2 samples. However, results from different workflows showed different estimations of sizes of richness changes. Both the ultra-fast and fast workflows estimated that the richness of functional genes were ~12% lower in L3 samples than in L1 samples, the moderate, sensitive and ultra-sensitive workflows estimated that the richness of functional genes were ~24% to ~25% lower, and the SEED Subsystem annotation estimated that it was only ~2.8% lower. Meanwhile, the fast workflow estimated that that richness of functional genes was ~8.4% lower in L3 samples than in L2 samples, and SEED Subsystem annotation estimated ~2.3% of lower richness.

The above results on both composition and richness of functional genes indicated that the impacts of contaminants on groundwater metagenomes were detectable in the analyses based on all workflows, but was reflected in higher magnitude (DCA separation and sizes of richness changes) in the moderate, sensitive and ultra-sensitive workflows. These results met our expectation, which were likely due to two reasons. First, the databases of the EcoFun-MAP are more specific in terms of ecological functions (e.g. Metal Homeostasis) that are susceptible to impacts of the contaminants than the SEED Subsystem database. Second, the workflows where HMM model based annotation was involved were likely to have higher precision than the workflows (the ultra-fast, fast and SEED Subsystem workflows) based solely on sequence identity searching, and consequently, less noise was included.

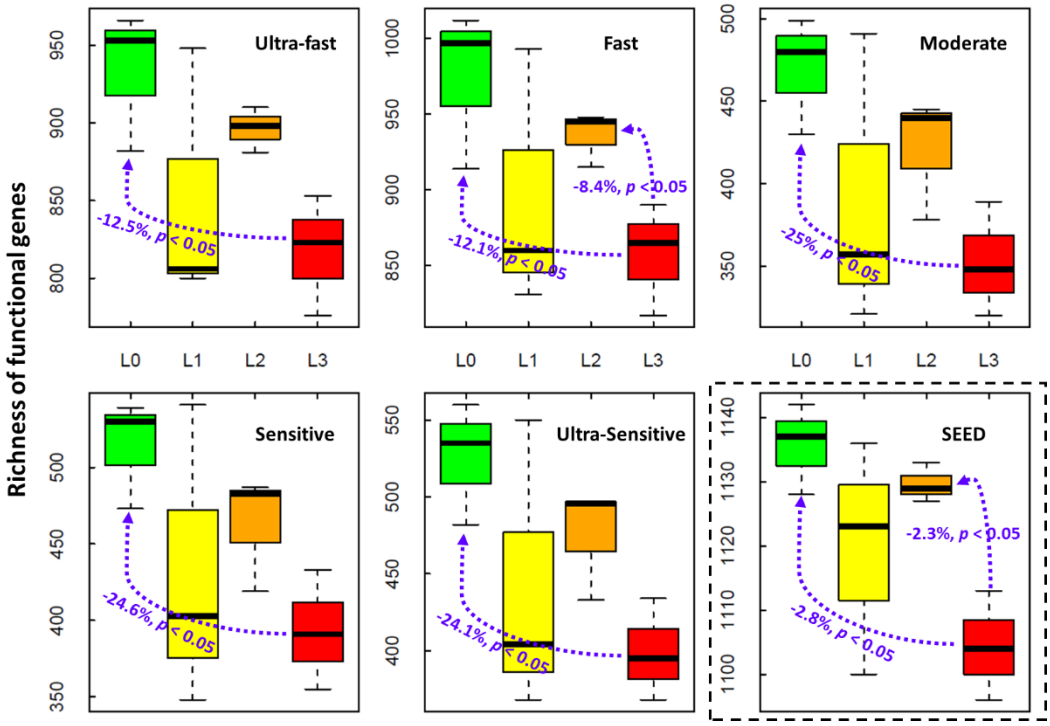


Figure 4.5 Richness of functional genes in metagenomes from 12 underground water samples. A total of six boxplots show the richness of functional genes based on results from five workflows of the EcoFun-MAP, as well as result from annotation based on SEED subsystem (boxed by dashed line). Boxes in color of green, yellow, orange and red are used for showing richness of functional genes for samples from groups of L0, L1, L2 and L3, respectively.

Next, we further analyzed relative abundances of major functional categories, including the category of C, N, S and P cycling, Metal homeostasis, Stress, Organic contaminant degradation, Antibiotic resistance, Electron transfer, and Virulence, which are considered highly relevant to the study site, and compared them among different samples. The analysis was based on the Ultra-sensitive workflow. Overall, relative abundances of functional genes from the C cycling category were lower in two of L3 samples (FW106 and FW021), which are two samples with highest level of contamination in many heavy metals (e.g. Cr, Eu and Ce) (**Figure S 9**), but those from the metal homeostasis category in the two samples were higher than other samples (**Figure 4.6**). Interestingly, sample FW104 from group L3, which had the highest level

of Sulfate (SO_4) of all samples (**Figure S 9**), also has the highest relative abundance of S cycling genes (**Figure 4.6**).

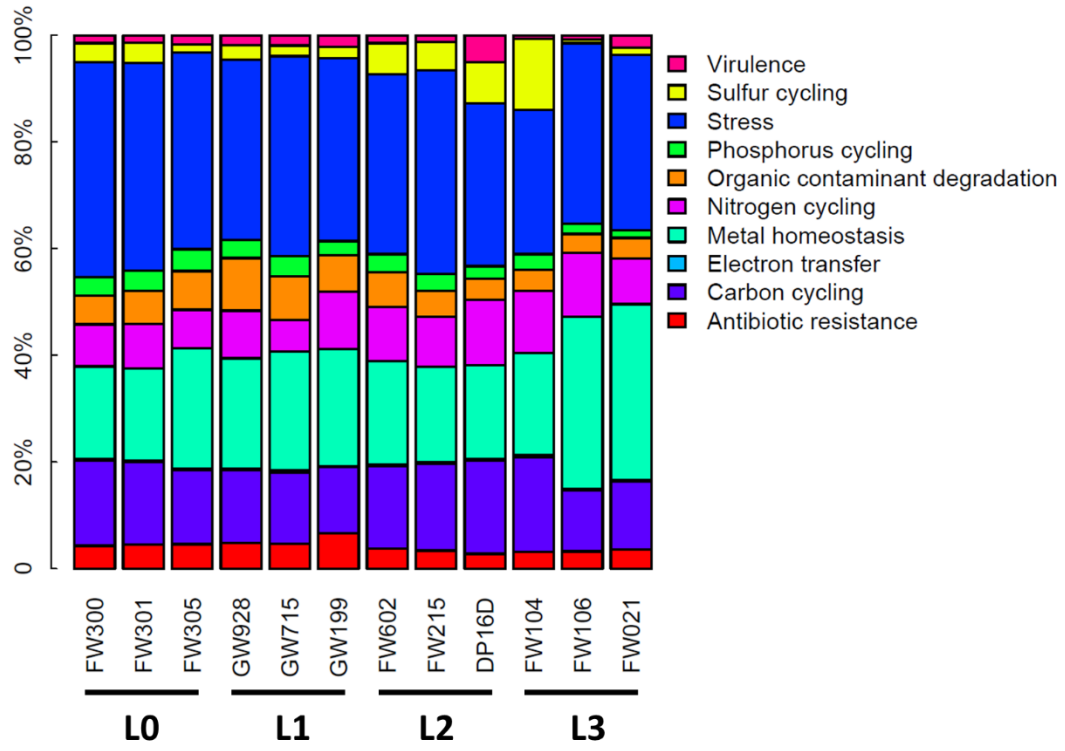


Figure 4.6 Relative abundances of selected major categories (based on result from Ultra-sensitive mode) in metagenomes from 12 underground water samples.

Response ratios of relative abundances of functional genes were calculated and compared between sample group L0 and each of other groups (L1, L2, and L3, **Figure 4.7**). We found significant positive response ratio of metal homeostasis genes in both L2 - L0 (*arrA* and *arxA*) and L3 - L0 (*corC*, *pcoA*, *mgtA*, and *merP*) comparisons, and significant negative response ratio of one C degradation gene (*ara*) in L3 - L0 comparison. A denitrification gene (*nirK*) had significant positive response ratio in L3 - L0 comparison, which suggested a microbial response to higher nitrate concentrations in the L3 samples (**Figure S 9**). Several oxygen-limitation-response genes (*narH* and *narJ*) from Stress category were more abundant in L3 samples than L0 samples, which

suggested a microbial response to low dissolved oxygen in highly contaminated wells (data not shown).

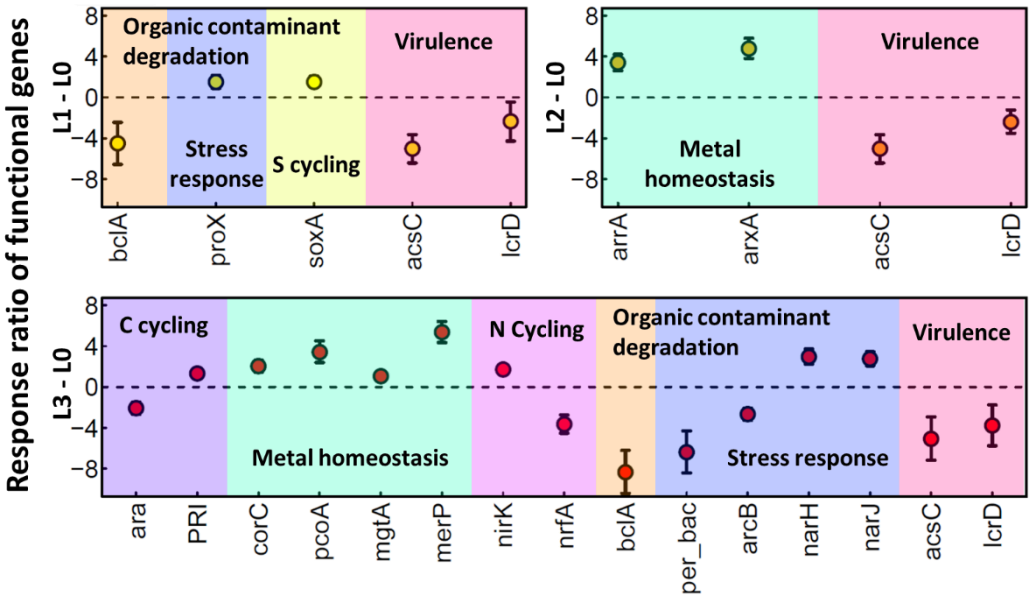


Figure 4.7 Response ratio of functional genes from comparisons between metagenomes from contaminated well samples and background well samples. Only significantly (*p value* < 0.05 in ANOVA followed by TukeyHSD) changed genes are included in the plot.

4.5 Discussion

We developed the EcoFun-MAP in this study, including database construction, workflow design and pipeline implementation and deployment. The EcoFun-MAP provides an efficient, flexible and accessible for analyzing high-throughput metagenomics sequencing data from an ecological function perspective. The EcoFun-MAP is thus capable of addressing some of the computational barriers brought by rapid throughput increase in NGS technology and faced by many microbial ecologists.

The databases of EcoFun-MAP have been constructed in this study and have several unique features and advantages. First, the databases have a selective coverage of functional genes that are important to microbial ecology studies. The databases are smaller and less in redundancy than databases that are more general, but still have the

comprehensive coverage for the defined scope that should enable effective functional analysis of metagenomics sequencing data. Also, the coverage matches up with the GeoChip, whose coverage has been demonstrated to be effective in numerous real world microbial ecology studies. Second, the quality of reference sequences was ensured using two separate procedures with manual corrections: the keyword query search and iterative confirmation using HMM of seeds, thus the reference sequences used for database construction should be accurate. Third, the reference sequences of each functional gene were clustered into fClusters, which provide resolution higher than gene and thus allow analysis that is more detailed. In addition, the databases of EcoFun-MAP were offered in a variety of widely accepted data structures, including indexed protein sequences (EFM-DI-DB-S and EFM-DI-DB-R), HMM models (EFM-HMM-DB), and indexed nucleotide sequences (EFM-BLAST-DB), which not only allows different levels of speed and sensitivity in analysis, but also provide multiple interfaces for potential future extension if there will be new tools having better annotating algorithms but relying on the same data structure. All of above features should make the databases in the EcoFun-MAP valid in analyzing metagenomics sequencing reads from the perspective of ecological functions.

The EcoFun-MAP is open for public use in a form of website, so it is free of installation and configuration of software or databases, and can be accessed using plain web browsers easily with an Internet connection. While the EcoFun-MAP was implemented and deployed based on sophisticated hardware and bioinformatics tools, it requires little computational skills to use other than simple web-based user registration, uploading of datasets, and mode selection or parameter setting. Furthermore, the EcoFun-MAP has

multiple predefined workflows built on disparate databases and tools, so it provides a flexibility for addressing different needs for speed or sensitivity at little cost of ease of use. Finally, the EcoFun-MAP is supported by an HPC infrastructure, which provides access to advanced hardware resources (e.g. fast CPUs, large memory and hard disk space) required by data-intensive projects for public use. Therefore, the EcoFun-MAP should be easy for microbial ecologists to access and use.

With a typical speed of analysis from ~0.6 to ~4.1M reads/min. (highest speed for workflows), the EcoFun-MAP is considered to have the desirable speed for metagenomics sequencing data analysis in microbial ecology, especially for handling large (>10Gbp) dataset. Due to a lack of availability and speed reports of functional analysis pipelines, and differences in the configuration of hardware and software, fair speed comparisons between the EcoFun-MAP with other pipelines were difficult to make. To our best knowledge, the EcoFun-MAP is the first web-based pipeline with speed of multi-million reads per minute. The EcoFun-MAP gained speed advantages through several features. First, smaller reference databases with a clear focus, cleaned and optimized for selectively annotating reads with information of functional genes are that important or highly relevant to ecological functions or geochemical process. With our curation effort, the EcoFun-MAP databases only have 1.5% of the size of NCBI RefSeq database (81,027,309 protein sequences; release 81, Mar 13th, 2017). Such reduction strategy has been shown as a practical solution for speeding up high-throughput sequencing data analysis (Silva, Green et al. 2016). Second, fast and updated tools were selected for the EcoFun-MAP and contributed substantially to the fast speed of the EcoFun-MAP. For example, FragGeneScan+ used for gene prediction is 5-50

times faster than FragGeneScan at basically no cost of performance in terms of other aspects (e.g. accuracy) (Kim, Hahn et al. 2015). HMMER 3 is 100-1000 times faster than HMMER 2 (Eddy 2011). DIAMOND is 20,000 times faster than BLASTX (Buchfink, Xie et al. 2015). Third, the EcoFun-MAP was implemented with parallel processing feature and deployed on HPC clusters, which gains additional acceleration from a hardware perspective. In addition to the speed itself, the EcoFun-MAP is also highly scalable, which is quite important because the volume of sequencing data is well expected to increase in the foreseeable future. The EcoFun-MAP will then still be able to contain the time cost of the analysis.

Accuracy and precision of pipelines for analyzing metagenomics reads are important but difficult to evaluate. Both of the accuracy and precision will change if parameters for processing steps are adjusted. However, we found in several ways that the EcoFun-MAP should have adequate performance for functional analysis of shotgun metagenomics sequencing data. First, the reference databases of the EcoFun-MAP were accurate as previously discussed, which provided foundations for quality of analysis. Second, the EcoFun-MAP was designed to rely on protein sequence based searches rather than the ones based on nucleotide sequence, which are considered to be more accurate. Furthermore, HMM searches are also used in three workflows of the EcoFun-MAP, which are more time-consuming but more accurate than read identity based searches. In addition, the EcoFun-MAP provides 5 predefined workflows which are different in accuracy and precision. In general, the workflows provided accuracy rates more than ~70%, but differed in terms of precision rates, which are much lower in the ultra-fast and fast workflows. It is likely because that the ultra-fast and fast workflows

are dependent on the read identity based searches, which are more likely to introduce unreliable hits more capable of discovering novel fragments of target gene. The variations in accuracy and precision in the predefined workflows are helpful for addressing different needs in real studies. For example, the ultra-fast and fast workflows may be inappropriate for studies requiring strict control on false positives, but should be competent for studies with explorative emphasis. The sensitive workflow had both high accuracy (> ~85%) and precision rate (> ~85%), as well as speed (1.2M reads/min. in average) in our analysis, thus was set to default mode for the EcoFun-MAP. Furthermore, all EcoFun-MAP workflows were capable of generating similar results in analyses of composition and functional gene richness of the metagenomes from underground water samples in our analysis. Detailed analyses of the metagenomes based on the EcoFun-MAP demonstrated its usefulness in revealing differences in relative abundances of functional categories and functional genes among sampled microbial communities, and link the differences with different levels of contaminants in the samples.

The EcoFun-MAP, at its first version, has several limitations. First, the coverage of the EcoFun-MAP is by no means complete, though it is comprehensive in the scope defined by this study. Some ecological functions were less understood and key genes involved these functions are not known, thus it is impossible to include them. Meanwhile, a vast majority of diversity in reference sequences for some genes was lacked in the data source of the EcoFun-MAP (NCBI databases), thus the coverage for these genes is also incomplete. The limited coverage may cause drops in sensitivity when analyzing the sequencing datasets. Second, the quantitative capability of the EcoFun-MAP is limited,

though as a pipeline handling unassembled reads, it preserves more critical frequency information than the methods dependent on assembly. This limitation is common in currently available pipelines, because revealing true information about which genome and what position each read is from is still challenging due to the confidence of imperfect matches between reads and reference sequences, biases of bioinformatics tools, and the limited coverage. When the information is absent, converting the read frequencies to gene abundances will be inevitably biased. Third, the EcoFun-MAP is not appropriate for obtaining accurate taxonomical/phylogenetic profiles from shotgun metagenomic sequencing datasets, because coding sequences are not good references for identifying taxonomical/phylogenetic units. It not rare at all that different microbial species can have highly similar or identical coding sequences for a same functional gene. To lift above limitations, it is important to keep updating the EcoFun-MAP by incorporating emerging reference sequences of genes of interest, upgrading or replacing bioinformatics tools, and building independent databases of phylogenetic markers. Our other future plans include adding modules for data visualization and downstream comparative analysis, and upgrading hardware for future hikes of data size.

4.6 Conclusion and availability

In this study, we developed the EcoFun-MAP for functional analysis of shotgun metagenomics sequencing data from microbial ecology. The EcoFun-MAP consists of references databases constructed with selective coverage of genes that are important to ecological functions, and multiple workflows for addressing disparate needs for speed and accuracy. Furthermore, the EcoFun-MAP was implemented on the basis of High-Performance Computing (HPC) infrastructure with high accessible interfaces. In our

analysis, we found the EcoFun-MAP a fast and useful pipeline for functionally profiling metagenomes from underground water samples. The EcoFun-MAP is open for public use and can be found available at our website: http://zhoulab5.rccc.ou.edu:7999.

Chapter 5: A generalized Brody distribution based Random Matrix Theory approach for inferring microbial data association networks

5.1 Abstract

Microorganisms are not isolated from each other but rather receive and impose positive, negative or negligible real-time impact from and to each involved species. Therefore, identifying and investigating these interactions within microbial communities will not only help us to understand microbial responses to perturbation, but also ultimately improve the predictive capability of global models of ecosystem dynamics. Because of its straight-forward calculation procedure and high processing speed, data association network inference has become a widely-adopted approach for efficiently inferring networks from large and complex microbial community systems. Inference of data association networks relies on a crucial step where a critical threshold is chosen for removing links with association strength below the threshold. Most studies have selected thresholds empirically or arbitrarily, thus the inferred networks are inevitably susceptible to biases and lead to inaccurate inference and analysis of networks. We previously proposed a Random Matrix Theory (RMT)-based approach for detecting objective thresholds automatically, but it still had limitations in terms of capability of detection and interpretability of detected thresholds. Here we developed a new method based on the generalized Brody distribution (GBD) for determining the critical threshold in the framework of the RMT, and proposed an improved approach (GBD-RMT approach) for inferring microbial data association networks. Results showed that the GBD-RMT approach is capable of quantitatively characterizing the dynamics of Nearest Neighboring Spacing Distribution (NNSD) of eigenvalues against candidate

thresholds, and detecting both the critical transitions and thresholds in NNSD dynamics using trend analysis. In our evaluation, the GBD-RMT approach successfully detected the critical thresholds in all of the numerically simulated and real datasets, including those for which the previous method failed. It also had higher detection resolution, and gained higher confidence and interpretability in detected critical thresholds. Meanwhile, the GBD-RMT approach integrated improvements for detecting more types of data association and reducing compositional data bias. In addition, the GBD-RMT approach uncovered a remarkable overlap between the critical transitions and the plateaus of scale-freeness from the inferred networks, and the overlap is showed to be statistically significant and universal in complex biological systems in our analysis. In conclusion, the GBD-RMT approach proposed in this study presented a powerful and state-of-art tool in inferring microbial data association networks.

5.2 Introduction

The microbial communities are complex in organization and dynamics of interactions among microorganisms. Through the interactions, microbes as a community can orchestrate system level functions or exhibit influential properties which are impossible to study based on the disconnected populations. Therefore, the rudimentary taxonomical and functional profiling and comparative analysis is not sufficient for understanding and predicting microbial communities, and it now has become increasingly important to characterize microbial interactomes, especially for advancing our understanding of microbial diversity (Deng, Jiang et al. 2012), species co-evolution (Shi, Nuccio et al. 2016), microbial responses to the perturbation effects, and in microorganisms into predictive models of ecosystem dynamics (Shi, Nuccio et al. 2016). Network analysis has been widely adopted in diverse studies of complex species interactions in macro systems. For example, the organization food webs (Dunne, Williams et al. 2002) and pollination networks (Kaiser‐Bunbury, Muff et al. 2010) have been demonstrated to be linked to system persistence and disturbance, or to species coexistence and diversity. Yet, the network analysis, especially the network inference, in microbial community studies is more challenging, due to the extreme diversity of microorganisms and a consequent lack of complete accurate maps of interactions on the basis of biological knowledge. Therefore, it became increasingly important to infer the community interactions directly from empirical profiling data without prior knowledge.

Given quantitative abundance data for each component in a microbial community, we can apply a group of network inference methods to predict the interactions among the components without additional information or prior knowledge. The group of network

inference methods can be divided into two major types, multiple regression based inference and pairwise data association based inference (Faust and Raes 2012). The first type of methods models the abundance of one component as a function of all other components with corresponding optimized coefficients, and the second type estimates an association strength signaled by the correlation/dependence/co-occurrence pattern that can be detected between the abundances of two components. The network inference procedures developed based on the principles from two types of method can be very different and have distinctive strengths and disadvantages, and they were also demonstrated to be complementary to each other (Faust and Raes 2012, Faust, Sathirapongsasuti et al. 2012). Although the first type of method suits better for recovering more complex interactions (i.e. one component is co-affected by more than one components), its capability is usually limited in underdetermined systems where sample number n is smaller than variable number p. Some assumptions were made to ease the problem, such as a positive linear relationship between the number of links and the number of components (aka. linear sparsity), but these assumptions are not always based on biological reasons, don't necessarily hold for every complex system, and sometimes complicate the interpretation of regression results. The second type of methods is conceptually straightforward, computationally simple and parallelizable, and less constrained by the dimensionality problem, thus is most commonly used for inferring networks from high-throughput sequencing and microarray data (typically highly dimensional and significantly under-sampled) in soils (Shi, Nuccio et al. 2016), oceans (Steele, Countway et al. 2011, Lima-Mendez, Faust et al. 2015), lakes (Eiler,

Heinrich et al. 2012), and even in global genomic surveys (Lima-Mendez, Faust et al. 2015).

However, caution has to be used when inferring data association networks. First, spurious associations can be introduced due to the simplex nature of compositional data (Friedman and Alm 2012), since relative abundance data from high-throughput technologies was prevalently used in microbial ecology studies. The problem can be alleviated using several data preprocessing techniques previously reported, including Compositionally Corrected by REnormalization and PErmutation (CCREPE) (Faust and Raes 2012, Faust, Sathirapongsasuti et al. 2012) and Aitchison's transformation based method (Friedman and Alm 2012, Kurtz, Müller et al. 2015). Second, popular data association methods for inferring networks were the Pearson Correlation Coefficient (PCC) and Spearman Rank Coefficient (SPM), which are limited in types of data association that are detectable. Both the methods are not appropriate for detecting many non-linear or non-functional data associations. Third, because inference of data association networks relies on a crucial step in which a critical threshold is selected for deleting links with association strength less than the threshold, using inappropriate critical thresholds can cause the inaccurate structure of inferred network or difficulty to interpret. Using arbitrary thresholds, or selecting thresholds based on empirical p-values or the optimization over designated topological properties (e.g. scale-freeness) tend to introduce bias in network inference or interpretability issue. To address this problem, we previously presented an approach (Luo, Zhong et al. 2006, Luo, Zhong et al. 2006, Zhou, Deng et al. 2010, Zhou, Deng et al. 2011) based on the Random Matrix Theory

(RMT) (Mehta 2004), which is able to automatically and objectively identify a critical threshold.

The RMT-based threshold detection has several advantages. First, the method approach is developed based on the two universal laws of the RMT, and thus it is theoretically sound. Second, the threshold detection in the RMT-based approach is automatic and objective. Third, since RMT is powerful for removing noise from nonrandom, system-specific features, the inferred network is reliable. Fourth, the applicability of the RMT in biological systems has been demonstrated for inferring metabolic, protein, functional gene and microbial ecological networks (Luo, Yang et al. 2007, Luo, Yang et al. 2007, Zhou, Deng et al. 2010). However, the current RMT approach, the MENAP, have several limitations. First, the MENAP is limited in detecting data associations other than linear correlation, as it relies on the PCC. Second, the MENAP doesn't have any preprocessing step to remove compositional data bias. Third, the MENAP failed to detect critical thresholds occasionally. In addition, the MENAP calls critical threshold on each candidate cutoff without telling how good it is for the threshold, which is a lack of quantitative assessment of transition progress, and made the inferred networks less interpretable.

Thus, in this study, we will provide a generalized Brody distribution (GBD) based Random Matrix Theory approach (GBD-RMT approach) for inferring microbial data association networks. The GBD-RMT approach acquires the GBD unifying Wigner-Dyson and Poisson distribution with one single parameter, β, that can be used as a quantitative indicator of the transition progress of the NNSD. Maximum Likelihood Estimation (MLE) based method was used for obtaining a best estimation for the β.

Meanwhile, the critical transitions and thresholds were detected using trend analysis on the β dynamics generated from the snapshots of a series of data association matrix reductions with cutoff values from low to high. In the evaluation of the GBD-RMT approach, both *in silico* and real datasets were used for demonstrating the effectiveness of the approach. Comparisons were also made between the GBD-RMT approach and the previous approach (Luo, Zhong et al. 2006, Luo, Yang et al. 2007) to show the advantages of the GBD-RMT approach. In addition, with the GBD-RMT approach, we uncovered a remarkable overlap between the critical transitions of the β dynamics and the plateaus of scale-freeness from the inferred networks, and the overlap is showed to be statistically significant and universal in complex biological systems in our analysis.

5.3 Materials and methods

5.3.1 Preprocessing of compositional data

Compositional data, in forms of fraction, proportion or relative abundance data, is commonly generated and used in microbial ecology studies, due to nature of high-throughput technologies, sampling and resampling methods or data normalization and transformation. Compositional data bears only relative information about its components, which are usually non-negative real values and sum up to a constant. Given n components and m samples, a typical compositional data, C, then can be defined as the following,

$$C = \begin{pmatrix} C_{11} & \cdots & C_{1m} \\ \vdots & \ddots & \vdots \\ C_{n1} & \cdots & C_{nm} \end{pmatrix}$$

and

$$\sum_{i=1}^{n} C_{ij} = k; j = 1,2, \dots, m;\ C_{ij} \geq 0$$

where C_{ij} is the relative information of component i in sample j, and k is a constant (e.g. 1). Compositional data like the C share many common properties, and most importantly, it should not be directly used for further computation or analysis based on absolute abundances, such as correlation (i.e. the PCC) calculation. Therefore, preprocessing is always recommended to transform C into the original or other appropriate sample space first before applying analysis required absolute information (Reimann and Filzmoser 2000, Filzmoser and Hron 2009). Aitchison first introduced several transformation techniques on the basis of log-ratio of compositional data, especially the centered log-ration transformation (clr), to transform compositional data to an unconstrained real space (Aitchison 1986). The clr transformation is defined as the following:

$$clr(C_{*j}) = \left[\ln\frac{C_{1j}}{g(C_{*j})}, \ln\frac{C_{2j}}{g(C_{*j})}, \dots, \ln\frac{C_{ij}}{g(C_{*j})}, \dots, \ln\frac{C_{nj}}{g(C_{*j})},\right] = \ln\frac{C_{*j}}{g(C_{*j})}$$

and

$$i = 1,2,\dots,n; j = 1,2,\dots,m;$$

where C_{*j} is abundances of all components in sample j, and $g(C_{*j})$ is the geometric mean of the C_{*j}, i.e.

$$g(C_{*j}) = \sqrt[n]{\prod_{i=1}^{n} C_{ij}}$$

The clr transformed variables can be interpreted as the original variables, but correlations between them cannot be interpreted in the same way (Filzmoser and Hron 2009). Still, in general, there is also no way to transform the correlations back to the original space (Filzmoser and Hron 2009). Approximately, the correlations between clr

transformed variables are shown to be equal to the basis correlations (i.e. correlations in original space), under an assumption that the C has a large number of components which are only sparsely correlated (Friedman and Alm 2012). In fact, the assumption holds for most microbial ecological data, thus here we also use the clr transformation for preprocessing compositional data before applying correlation calculation. When being applied to compositional data, calculation of other data associations than correlation can also be biased, but likely in different ways. So, it will not be mathematically sound to recruit the same clr transformation for calculating other data associations, and specific transformation technique can be necessary for each of them. Unfortunately, such transformation techniques have not been established yet, therefore the compositional data preprocessing described here was only used for correlation calculation. Other data associations were calculated directly without the preprocessing.

5.3.2 Calculation of data association matrix

The term data association, equivalent to dependence, is defined as any interesting relationship between two random variables that not satisfy probabilistic independence. It includes correlation (i.e. linear dependence), non-linear or non-functional dependence. The data association strength between any two microbial taxonomical or functional units can be estimated from the changes in their abundances in various samples or biological replicates over time or space. Therefore, a matrix consisting of data association strengths from all possible pairs of taxonomical or functional units can be computed from typical microbial ecological dataset:

$$P = \begin{pmatrix} P_{11} & \cdots & P_{1n} \\ \vdots & \ddots & \vdots \\ P_{n1} & \cdots & P_{nn} \end{pmatrix}$$

and

$$P_{ij} = D(C_{i*}, C_{j*}) = P_{ji}$$

where P_{ij} is the data association strength between component i and j, C_{i*} is abundances of component i in all samples, D denote any given method for measuring data association strength between C_{i*} and C_{j*}, and n is the total number of components. There are many methods available to be to be given as D, and we incorporated several types in this study to not fully compare their features or properties in characterizing microbial data association but to demonstrate that different data association methods can affect network inference and end networks. Each type was listed and briefly described below.

a) PCC and CLR_PCC. Karl Pearson came up with the PCC to measure the linear association between two variables (i.e. C_{i*} and C_{j*} here) in the 1880s, which is popular and widely used in many disciplines. The term correlation is sometimes used to refer to the PCC, if specific indication was absent. The PCC is calculated by dividing the covariance of the two variables by the product of their standard deviations, i.e.

$$PCC_{ij} = \frac{E[(C_{i*} - \mu_{C_{i*}})(C_{j*} - \mu_{C_{j*}})]}{\sigma_{C_{i*}}\sigma_{C_{j*}}}$$

and

$$i, j = 1,2, \dots, n;$$

where E is the expectation, μ is the mean and σ is the standard deviation. The PCC have a value range from -1 to 1, where -1 and 1 mean perfect negative and positive linear association, and 0 means no linear association. The CLR_PCC is simply the PCC between clr transformed variables.

b) Spearman's Rank Correlation Coefficient (SPM). The SPM is another popular method, which was invented by Charles Spearman to measure the correlation between the rankings of two variables. The SPM can be calculated by substituting the variable values with variable value ranks in the PCC formula, i.e.

$$\mathrm{SPM}_{ij} = \frac{E[(rank(C_{i*}) - \mu_{rank(C_{i*})})(rank(C_{j*}) - \mu_{rank(C_{j*})})]}{\sigma_{rank(C_{i*})}\sigma_{rank(C_{j*})}}$$

Regardless of linearity of data association, the SPM assesses monotonicity. If values in C_{i*} increase monotonically as values in C_{j*}increase, the SRCC is 1; if values in C_{i*} increase monotonically as values in C_{j*} decrease, the SPM is -1; if values in C_{i*} changes independently from changes of values in C_{j*}, the SPM is 0.

c) Kendall rank correlation coefficient (KDL). The KDL also measures correlations of ranks like the SPM, but it does not use the PCC formula. The KDL can be calculated by dividing the difference between number of concordant pairs and number of discordant pairs in two joint random variables by total number of pair combinations, i.e.

$$\mathrm{KDL}_{ij} = \frac{\sum_{k=1}^{m}\sum_{l=k+1}^{m} concor(C_{ik}, C_{jk}, C_{il}, C_{jl}) - \sum_{k=1}^{m}\sum_{l=k+1}^{m} discor(C_{ik}, C_{jk}, C_{il}, C_{jl})}{m(m-1)/2}$$

and

$$concor\left(C_{ik}, C_{jk}, C_{il}, C_{jl}\right) = \begin{cases} 1, & C_{ik} > C_{il} \text{ and } C_{jk} > C_{jl} \\ 1, & C_{ik} < C_{il} \text{ and } C_{jk} < C_{jl}; \\ 0, & else \end{cases}$$

$$discor\left(C_{ik}, C_{jk}, C_{il}, C_{jl}\right) = \begin{cases} 1, & C_{ik} > C_{il} \text{ and } C_{jk} < C_{jl} \\ 1, & C_{ik} < C_{il} \text{ and } C_{jk} > C_{jl}; \\ 0, & else \end{cases}$$

where m is the number of samples.

d) Distance Correlation (dCor). The dCor is another type of data association extending the PCC in detecting more complex associations between two variables other than linear dependence. The calculation of the dCor is based on the joint characteristic function and marginal characteristic functions in a weighted space, and the key advantage of the dCor is that it gives a value of 0 if and only if two random variables are independent, which is not guaranteed in the PCC. The details about the dCor can be found in (Székely, Rizzo et al. 2007).

e) Local Similarity Score (LSS). The LSS was introduced in local similarity analysis (LSA)(Ruan, Dutta et al. 2006), which is another measure aiming at identifying more complex data associations than the PCC. The LSS is selected to be the maximal sum of the product of all possible subvectors of two random variables within some predefined time delay D, which is not applicable in our analysis. The local similarity score is computed by dynamic programming. The computing procedure of LSS is described in details elsewhere (Ruan, Dutta et al. 2006).

f) Mutual Information (MI). The MI of two random variables is a measure of the mutual dependence between the two variables based on entropy. More specifically, the MI measures the similarity between the joint distribution of two random variables and the products of their marginal distributions, i.e.

$$\mathrm{MI}(C_{i*}, C_{j*}) = \sum_{k=1}^{m} \sum_{l=1}^{m} p(C_{ik}, C_{jl}) \log\left(\frac{p(C_{ik}, C_{jl})}{p(C_{ik})p(C_{jl})}\right)$$

where $p(C_{ik}, C_{jl})$ is joint probability distribution function, and $p(C_{ik})$ and $p(C_{jl})$ are marginal probability distribution functions.

g) Maximum Information Coefficient (MIC). The MIC is an MI based measure of data association. Since real world MI calculation can depend on a binning scheme, choosing

different numbers of bins and layouts of binning grids might lead to different final MI values. The MIC is selected as a value of normalized MI between two random variables that is maximized by searching the optimum number of bins and layout of grids with a heuristic algorithm. As a result, the MIC was claimed to preserve a property called equitability when measuring data association between any given two random variables, regardless of linearity and functionality of the data association. Being normalized MI values, thus the MIC values will always fall between 0 and 1. The calculation of the MIC is documented elsewhere in details in (Reshef, Reshef et al. 2011)

Among all the methods, the values of entries in the P of the CLR_PCC, LSS and MI were normalized by dividing the absolute value of the maximum P_{ij}.

5.3.3 The RMT approach framework

The aforementioned data association matrix, P, are influenced by "noise", so the crucial process here is to separate noisy or random constitutes of the P from the true ones. Here, an RMT-based framework is used for cleaning the P. The central assumption of RMT-based framework is that any given data association matrix like the P should consist of both random noise and system specific properties, and they can be distinguished because the noise should have weaker strength than non-random co-occurrences. Based on the assumption, the network inference can be transformed into a problem finding a critical threshold which is higher than most if not all noise and lower than the true co-occurrence associations. Therefore, if we define a function,

$$F(P, s): P \xrightarrow{s} R,\ T, \forall\ 0 \leq s \leq 1,$$

to capture a process dividing P into R and T with s in a way such that,

$$\begin{cases} R_{ij} = P_{ij},\ if\ |P_{ij}| < s \\ R_{ij} = 0\ \ ,\ if\ else \end{cases},$$

and

$$\begin{cases} T_{ij} = P_{ij},\ if\ |P_{ij}| \geq s \\ T_{ij} = 0\ \ ,\ if\ else \end{cases},$$

for

$$\forall\ 1 \leq i, j \leq n,$$

where s is the critical threshold, R is the matrix consisting of weaker co-occurrence associations, T is the matrix consisting of stronger ones, and both R and T have the same dimensions as P, then we can formulate the problem into searching for a single s resulting in an R and T from $F(P, s)$ such that,

$$\begin{cases} R \cong R_0 \\ T \cong T_0 \end{cases},$$

where R_0 is the presumptive matrix consisting of only noise or random co-occurrence associations, and T_0 is the presumptive matrix consisting of only true co-occurrence associations. Given any complex system P, there exists only one R_0 and T_0, and any one of R_0 and T_0 is calculable if the other is known. Therefore, the key in this framework for searching the best critical threshold is a mathematically solid reference for either R_0 or T_0, which is, however, elusive in most of the biological systems. Fortunately, such a reference point for T_0 has been shown to be approachable from eigen-spectra analysis in the RMT. The real symmetric matrix systems obey two universal laws in RMT, therefore if a T produced from $F(P, s)$ with some s has an NNSD of unfolded eigenvalues following Poisson statistics, its properties were indeed system specific and non-random. Previous studies showed that when s is low enough, resulted T is inflated

with weak randomness and noise, thus has an NNSD following Wigner-Dyson distribution:

$$PR_{Wigner-Dyson} \approx \frac{\pi}{2} \cdot d \cdot e^{(-\pi \cdot d^2/4)},$$

where d is the random variable referring to Nearest Neighboring Spacings (NNS) of unfolded eigenvalues; when s is high enough, the NNSD of T follows Poisson statistics:

$$PR_{poisson} \approx e^{-d};$$

when s was increasing from the lower to the higher, a transition of NNSD of T from following Wigner-Dyson to Poisson distribution was expected, and the corresponding T served as an approximation of T_0. This reference point was mathematically defined and could be automatically obtained, thus was considered to be objective.

5.3.4 The GBD-RMT approach

Apart from the previous approach, which testing whether the NNSD following Poisson or Wigner-Dyson distribution with χ^2 test (Luo, Yang et al. 2007, Zhou, Deng et al. 2010, Zhou, Wu et al. 2011, Deng, Jiang et al. 2012), our new proposed approach is to find the relationship between cut-off point (threshold value) and the parameter β of GBD (Sakhr and Nieminen 2006, Bandyopadhyay and Jalan 2007) which is used to describe the NNSD. The GBD used for describing the NNSD given by

$$PR_{Brody}(d) = (\beta + 1) \cdot \alpha \cdot d^{\beta} \cdot e^{(-\alpha \cdot d^{\beta+1})}$$

Where $\alpha = \left[\Gamma(\frac{\beta+2}{\beta+1})\right]^{\beta+1}$ and the parameter $0 \leq \beta \leq 1$. As $\beta = 0$, this distribution reduces to Poisson statistics, $PR_{poisson}(d) \approx e^{-d}$ where d is the spacing variable. As $\beta = 1$, this distribution goes to Wigner-Dyson statistics, $PR_{Wigner-Dyson}(d) \approx \frac{\pi}{2} \cdot d \cdot e^{(-\pi \cdot d^2/4)}$. Hence, the identification of transition of NNSD between Wigner-Dyson and

Poisson distributions in the original RMT based approach was converted into the estimation of the β parameter of GBD with NNS data. Therefore, we proposed the following general algorithm for identifying the transition of NNSD from empirical data,

1. Started from any given co-occurrence matrix, P, which is a $n \times n$ symmetric matrix, and every quantity p_{ij} from P is the co-occurrence strength (e.g. Pearson correlation) between the entity i and j from a total of n entities. The entity here may be equivalent to microbial species or taxonomic unit of other levels.
2. Set an initial threshold value, s_0, and generate a series of threshold values ranged from s_0 to maximum threshold value by small paces. The s_0 may have value ranged from 0 to 1, the maximum threshold value, s_{max}, is usually 1, and all paces have equal length l, which is usually less than 0.01. At the end of this step, a threshold series, $s = [s_0, s_1, \dots, s_k, \dots, s_{max}]$, was generated, where $s_k - s_{k-1} = l$.
3. For each s_k in every s, the following sub-procedure was performed for obtaining the parameter β_{s_k} of generalized Brody distribution
    i. Reduce the P was to P_{s_k} in such a way that any $p_{ij} \in [-s_k, s_k]$ was set to 0.
    ii. Calculate eigenvalues λ of the P_{s_k} from this equation $\left(P_{s_k} - \lambda I\right)v = 0$, where λ is the eigenvalues, v is the corresponding eigenvectors, and I is the identity matrix. Because P_{s_k} is symmetric, all $\lambda_i \in \lambda$ are real and λ can be sorted in an order that $\lambda_1 \leq \lambda_2 \leq \cdots \leq \lambda_n$.
    iii. Calculate unfolded eigenvalues e from λ with $e_i = N_{av}(\lambda_i)$, where N_{av} is the unfolding function of eigenvalues, which was obtained using the cubic spline interpolation on the original integrated density of eigenvalues. The unfolding step here is to remove the spurious effects due to the variations of spectral density.

iv. Calculate the NNS of unfolded eigenvalues, d, where $d_i = |e_{i+1} - e_i|$ for every $i = 1, 2, \dots, n-1$.

v. Estimate the parameter β_{s_k} of generalized Brody distribution from d using maximum likelihood method. The method details can be found in the next section of Materials and Methods.

At the end of this step, a generalized Brody distribution β parameter series, $\beta = [\beta_{s_0}, \beta_{s_1}, \dots, \beta_{s_k}, \dots, \beta_{s_{max}}]$, was obtained.

4. Identify the critical transitioning point from s_c from the s based on the trend of the β series, and the s_c will be chosen as the final critical threshold.

5.3.5 Maximum likelihood based β estimation

Maximum likelihood based method can be used for estimating the parameter β of generalized Brody distribution from the observed NNS of unfolded eigenvalues d, which was mention above. The ML-based method imposes more regular dynamics and less uncertainties in the estimated values than other methods (Jafarizadeh, Fouladi et al. 2012), particularly the Least Square Fit-based method, which is a well-known and widely used method, thus it was considered to be a more reliable tool for analyzing the fluctuation properties of the spectra of RMT systems (Jafarizadeh, Fouladi et al. 2012). The mechanism and properties of ML-based methods has been described extensively elsewhere (Scholz 1985), here we briefly describe how it was applied to the β parameter estimation in this study. The probability density function of generalized Brody distribution, $P_{Brody}(s|\beta)$, was described above. Suppose there was a set of observed NNS data points $d = [d_1, d_2, \dots, d_n\ ]$ that are independent and identically

distributed, then the likelihood function for the β parameter based on the observed d was defined as the following,

$$\mathcal{L}(\beta\,;\,d_1, d_2, \ldots, d_n) = f_{Brody}(d_1, d_2, \ldots, d_n \mid \beta) = \prod_{i=1}^{n} f_{Brody}(d_i \mid \beta)$$

If a maximum exists for the defined likelihood function, it is mathematically the same regardless of whether we maximize the likelihood or the log-likelihood function, because the log function increases monotonically. Therefore, we defined the corresponding log-likelihood function as the following for computational convenience,

$$\ln \mathcal{L}(\beta\,;\,d_1, d_2, \ldots, d_n) = \sum_{i=1}^{n} \ln f(d_i \mid \beta).$$

Further, we have maximum-likelihood estimator $\hat{\beta}_{mle}$ defined as,

$$\hat{\beta}_{mle} = \arg \max_{0 \le \beta \le 1} \ln \mathcal{L}(\beta\,;\,d_1, d_2, \ldots, d_n)$$

i.e., the set of β parameters that were box constrained between 0 and 1 and maximize the log-likelihood function, $\ln \mathcal{L}(\beta\,;\,d_1, d_2, \ldots, d_n)$.

Note that the $\hat{\beta}_{mle}$ should be a value chosen from 0 to 1, so a limited-memory modification of the BFGS quasi-Newton method allowing both lower and upper bounds [Byrd *et. al.* (1995)] was used for the maximization of the log-likelihood function.

5.3.6 Identification of critical transition and selection of final threshold

Note that the generalized Brody distribution based approach estimates a series of β parameters, $[\beta_{s_0}, \beta_{s_1}, \ldots, \beta_{s_k}, \ldots, \beta_{s_{max}}]$, for the threshold series, $s = [s_0, s_1, \ldots, s_k, \ldots, s_{max}]$. The subsequent step is to identify the critical transitioning point in the s as signaled by the trend and fluctuation in the β series. It is important in this step to employ an algorithm that is capable of not only automatically recognizing the transitioning phase of the β dynamics in different systems, where the estimated β's

quickly drop from values near 1 to 0, but also objectively selecting critical points from such transitions to ensure that the further constructed networks can be compared with each other on a similar basis. Here we adopted the smoothing and detrending techniques from trend analysis, accompanied by several critical transitioning indicators in complex systems, including trend slope, lag-1 autocorrelation, variance, and skewness, for jointly determining the critical transitioning point in this study. Thus, the detailed algorithm is described as the following,

1. Smooth the β using a Gaussian kernel smoothing function with bandwidths chosen separately for each specific case, so that the major trend of the β was kept without overfitting. For most of the analysis in this study, the bandwidth was set to a tenth of the span of the β.
2. Subtract the smoothed values from the β and obtain the remaining residuals. This technique is also called detrending, which removes the long-term trend from the original β and achieves stationarity.
3. Evaluate the series of detrended β residuals from its tail, and identify a critical point if a sharp transition exists in the original β. Since the smoothing bandwidth was usually larger than the span of the sharp transition, a lagging effect will be shown on the tail of smoothing line, i.e. the smoothing line drops slower than the original signals. So, the bifurcating point of the smoothing line and original β dynamics was used as a presumptive transitioning point, which is usually at the point with the largest slope on the trend line, and the slope m at s_k for any possible k was calculated as following,

$$m_{s_k} = \frac{\Delta\beta}{\Delta s},$$

where $\Delta s = s_{k+c} - s_k$ and $\Delta\beta = \beta_{s_{k+c}} - \beta_{s_k}$, all c in this study used a small value of 4. The critical transition was estimated to be the longest consecutive region of points with slope values larger than 0.13, and the critical threshold was select at the point in the critical transition where the slope is maximum.

4. Confirm the presumptive critical point produced in the above steps, using lag-1 autocorrelation, variance, and skewness. To compute three indicators, a sliding window of fixed size up to the transition point was applied the detrended β residuals. Here, a sliding window of half the size of the time series. The detailed calculation was described as following,

   i. Lag-1 autocorrelation root-sum-of-squares (RSSQ). An autoregressive model of order 1 (AR1) was defined as following,

   $$\beta_{s_{k+1}} = a_1 \beta_{s_k} + \varepsilon_{s_k}.$$

   AR1 was fitted with windowed data points (assuming M points per window) by an ordinary least-squares (OLS) fitting method, and the RSSQ of AR1 was estimated as following,

   $$RSSQ_{s_k} = \sum_i^M \varepsilon_{s_k}^i.$$

   ii. Variance. The variance of windowed data points, δ_{s_k}, was estimated using the standard deviation.

   iii. Skewness. The skewness of windowed data points was estimated using the Pearson's moment coefficient of skewness, which was defined as following,

   $$\gamma_{s_k} = \frac{\frac{1}{M}\sum_i^M (\beta_{s_k}^i - \bar{\beta}_{s_k})^3}{\left[\frac{1}{M-1}\sum_i^M \left(\beta_{s_k}^i - \bar{\beta}_{s_k}\right)^2\right]^{3/2}},$$

   where $\bar{\beta}_{s_k}$ is the mean of the data points fallen within the window.

With m_{s_k}, $RSSQ_{s_k}$, δ_{s_k} and γ_{s_k} for any possible s_k, we normalized the value ranges of four indicators to $[0, 1]$. For m, $RSSQ$ and δ, min-max normalization was done separately, so each point of each indicator have a value close to 1 if its original value is close to the maximum one among all values of the indicator, and a value close to 0 if that is close to the minimum one; for γ, min-max normalization was done based on the distance of original absolute value to 1, so each point of γ should have a value close to 1 if its original absolute value is close to 1, and a value close to 0 if that is close to the most distant one from 1.

5.3.7 in silico datasets

Simulation procedure and rules. The goal of the simulation here was to obtain the artificial co-occurrence matrices whose individual strength values were distributed in designated patterns. Since any co-occurrence matrix produced by popular correlation calculation methods is symmetric, and has absolute values ranged from 0 to 1, the general procedure used in this study to simulate a $n \times n$ co-occurrence matrix has the following steps to ensure the rules to be meet,

1. Draw $n \times (n-1)/2$ random deviates from given probability density function to form an $n \times n$ upper triangular matrix U.
2. Normalize U by dividing each random deviate with the maximum absolute value of all, to ensure each has the value ranged from -1 to 1.
3. Get the lower triangular matrix L by transposing U, and obtain the simulated co-occurrence matrix $A = U + I + L$, where I is a $n \times n$ identity matrix, to ensure the symmetry.

The sizes of all simulated matrices in this study were set to 500×500 to facilitate the computation, unless else was specified.

Simulated co-occurrence matrices from common continuous distributions. To explore the generality of the GBD-RMT method, we analyzed datasets simulated from a number of common continuous distributions, including uniform, normal, log-normal, exponential, logistic, beta, gamma, and Weibull distribution. The parameter setting for each distribution was selected to have distinguished shapes of density function. The simulated dataset based on the normal distribution was chosen as the model system to represent *in silico* datasets for the purpose of demonstration in the results

5.3.8 Real project dataset

The real datasets were ideal than the ones numerically simulated, as they were found to have mixed, complex or irregular distributing patterns (**Figure S 10**). Therefore, real datasets below were used in this study.

a) The MENAP datasets. MENAP is an open-accessible pipeline at Institute for Environmental Genomics, which provides an implementation of current RMT-based approach to construct ecological association networks. It hosts more than 6,000 co-occurrence matrices (mainly based on Pearson Correlation Coefficient) that were calculated from the real 16S sequencing and gene expression datasets. In this study, a total of 500 co-occurrence matrices were extracted from MENAP database. Each dataset was selected randomly, with criteria having at least 500 qualified components that with valid values in more than two thirds of the samples. Selected co-occurrence matrices can be used as direct input for threshold detection using the GBD-RMT approach.

b) Biogeographic survey. The biogeographic survey datasets consist of 16S sequencing profiles of 126 metagenomes from soil core samples that were taken from the following six forests (21 for each): Niwot (NWT), Andrews (AND), Harvard (HFR), Coweeta (CWT), Luquillo (LUQ) and Barro Colorado Island (BCI). The selected sites provide variation in ecosystem type from boreal to tropical forest. More details about the datasets can be found in (Zhou, Deng et al. 2016) and **Table S 5**. The BCI dataset was chosen as the model system to represent real datasets for demonstration in the results

c) Plant succession. The plant succession datasets consist of 16S sequencing profiles of 288 metagenomes from rhizosphere and bulk soil samples from a greenhouse experiment (Shi, Nuccio et al. 2015). The samples were taken from 18 harvests at 10 time points (2 seasons and 5 time points for each season) during the succession of *Avena fatua*, in which 8 harvests were from rhizosphere soil (except for the first time point of each season) and 10 harvests were from bulk soil. More details about the datasets can be found in (Shi, Nuccio et al. 2016) and **Table S 6**.

5.3.9 Topological indices

a. Scale-freeness

A network is scale-free if its degree distribution fits a power law, which means most of nodes have low degree and only a few nodes have high degree. Mathematically, the fraction of nodes in a scale-free network that has degree of k, is denoted as $p(k)$, then

$$p(k) \sim k^{-c},$$

or

$$\log\big(p(k)\big) \sim -c \log(k),$$

where c is a constant. Upon the definition of the scale-free network, scale-freeness is defined as the goodness of fit of degree distribution to a power law. Given a vector of node degrees, the degree distribution can be approximated by binning the node degrees. Considering that the number of bins may affect the distribution approximation and the scale-freeness estimation, we make multiple attempts in calculating scale-freeness for a network with different numbers of bins ranged from 5 to 1/5 of node number, and only the maximum scale-freeness was selected. Assume $k = [k_1, k_2, \dots, k_i, \dots, k_n]$ is a vector of node degrees for a network of n nodes, in which k_i is the degree of node i, the calculation of scale-freeness of a network has the following steps:

1. Select a bin number b from 5 to n/5
2. Bin k with b, and obtain the degree distribution $p_b(k)$
3. Get $\log(p_b(k) + 1)$ and $\log(k + 1)$
4. Fit linear model with $\log(p_b(k) + 1)$ and $\log(k + 1)$, and record the R^2
5. Increase value of b by 1, and repeat step 2 to 4.
6. The highest R^2 of all is selected as the scale-freeness.

b. Other indices

Other topological properties of network included in this study are connected node number, edge number, average connectivity (average degree), average shortest path, average clustering coefficient, and modularity. Most calculations will be accomplished through the *igraph* (Csardi and Nepusz 2006) packages in the R project.

5.4 Results

5.4.1 Overview of the GBD-RMT approach

A microbial data association network is an implicative map of various biological interactions (e.g. predation, competition, and mutualism) between microbial species in complex microbiomes. In such a network, nodes are the microbial species or OTUs, and links are associations between microbial species abundances. To construct microbial data association networks, finding appropriate thresholds of data association strength for reducing the numbers of links is the key. The GBD-RMT approach developed in this study can overcome limitations of the current method (e.g. MENAP), providing a generalized and objective way to finding the thresholds based on the context of the RMT.

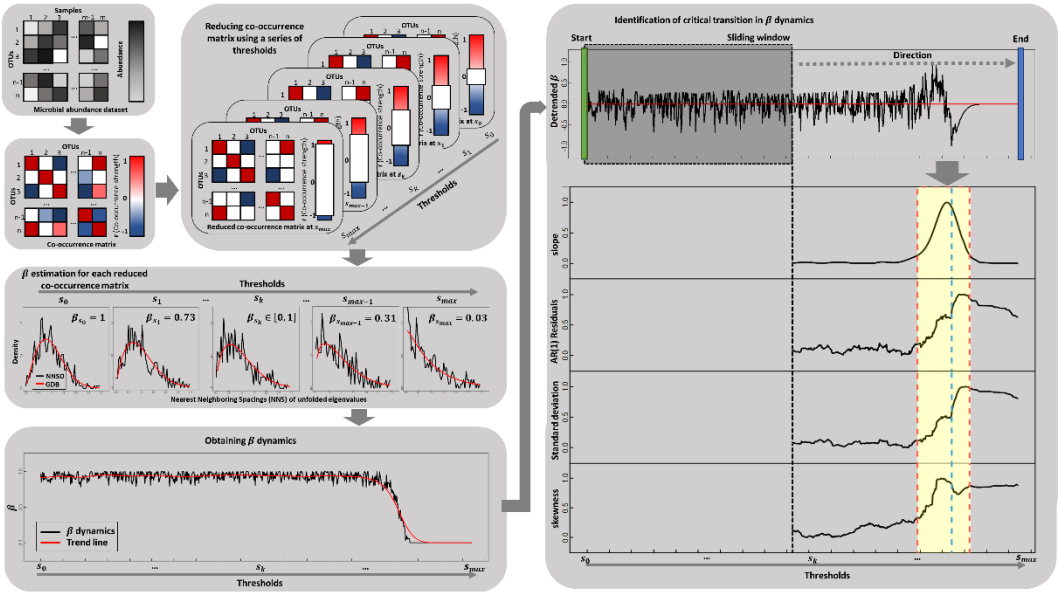


Figure 5.1 The schematic workflow of the GBD-RMT approach for determining critical threshold in datasets of species abundances.

The GBD-RMT approach here consists of 3 major steps (**Figure 5.1**). The first step is to calculate the data association matrices from abundance matrices. In this step, multiple (8) distinctive methods for estimating data association strength are included, which have

unique advantages and disadvantages for the calculation of data association matrix. When calculating correlation matrices (data association based on PCC) from compositional abundance data, the clr transformation was effective for reducing the biases imposed by the compositional nature. However, the clr transformation has not been shown to work for other data association estimation methods, and no other specific transformations were proved to be effective neither, thus the calculation of data association matrix from compositional data using the methods other than the PCC was enforced without corrections, and should be considered to have biases. The second step is to obtain β series for the data association matrix from the first step, and detect the critical transition and select a final threshold in the β series. With the finalized threshold, the data association matrix can be easily reduced to an adjacency matrix, which is a data structure for representing undirected networks. To this end, the network structure was fixed and the network was constructed. Third, network analysis is to be performed on the constructed network to investigate topological characteristics or properties of interest (e.g. scale-freeness), detect modules, identify keystone nodes, and link network properties to external factors (e.g. geochemical variables).

5.4.2 Generalized Brody Distribution

First of all, we verified the capability of the GBD in capturing the progressive shift from the Wigner-Dyson distribution to the Poisson distribution. The numerical simulation result (**Figure 5.2**) showed that when the value of parameter of β was equal to 1 (**Figure 5.2**a), the GBD had a probability density function (pdf) curve as the same as what the Wigner-Dyson distribution had (**Figure 5.2**b); when the β parameter was equal to 0 (**Figure 5.2**a), the GBD had a pdf curve that was the same as what the

Poisson distribution had (**Figure 5.2**c). When the value of the β parameter was deceased from 1 to 0 (1, 0.75, 0.5, 0.25 and 0) (**Figure 5.2**a), the GBD had transitioned from the Wigner-Dyson distribution to the Poisson distribution. When the value of the β parameter is closer to 1, the GBD is more like the Wigner-Dyson distribution; when the value of β parameter is closer to 1, the GBD is more like the Poisson distribution. The result suggested the GBD was indeed capable of capturing the transition between the Wigner-Dyson distribution and the Poisson distribution, if it existed.

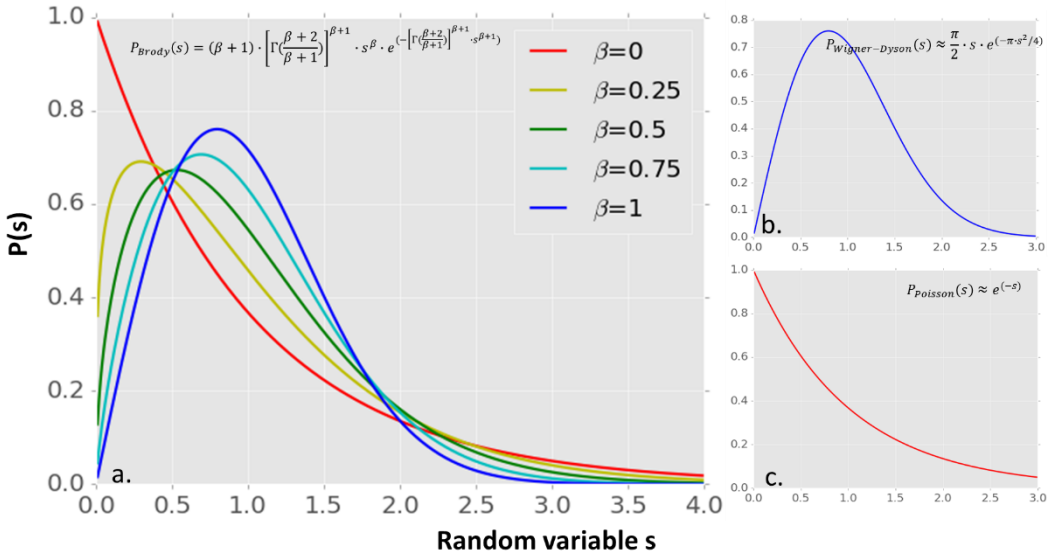


Figure 5.2 (a) The probability density function (pdf) of the Generalized Brody Distribution (GBD) with β values equal to 0, 0.25, 0.5, 0.75 and 1. (b) The pdf of Wigner-Dyson distribution. (c) The pdf of Poisson distribution.

5.4.3 Threshold detection in in silico datasets

The result in **Figure 5.3** showed an example of threshold detection with the GBD-RMT approach in a data association matrix (500 × 500) numerically simulated by randomly drawing strength values from a normal distribution ($\mu = 0$; $\sigma = 1$). The results showed that the critical transition of the NNSD in the simulated system existed and was successfully detected. The critical transition had a beginning point at ~0.81 and ending

point at ~0.91 and a critical threshold was selected at 0.867, which was the single point in the β series where the NNSD shifted in the fastest pace from the state that fit Wigner-Dyson distribution (β =1) to the other that fit Poisson distribution (β =0). The critical transition was also further evidenced by containing the peaks of AR (1) residuals, standard deviations, and skewness. The NNSD of the system was characterized by the GBD with a β value of ~ 0.5 at the selected final threshold, which suggested that it is close to a middle state that was well distinguished from both Wigner-Dyson and Poisson distribution. Similar results were also observed for the systems simulated by drawing from other continuous distributions, including exponential, log-normal, logistic, uniform, gamma, beta and Weibull distribution (**Figure S 11**). The critical transitions were found to exist in all the simulated systems, but had the different beginning (~0.4-0.9999) and ending points (0.49-0.99999), and the selected final thresholds were also ranged from ~0.44 to 0.99998. Interestingly, the critical transitions detected for the simulated systems can have very different spans, which were ranged from ~0.0001 (beta distribution) to ~0.1 (logistic distribution). These results indicated that the critical transitions and final thresholds existed in all the analyzed *in silico* datasets and were detectable to the GBD-RMT approach, but their value ranges can be distinctive from each other.

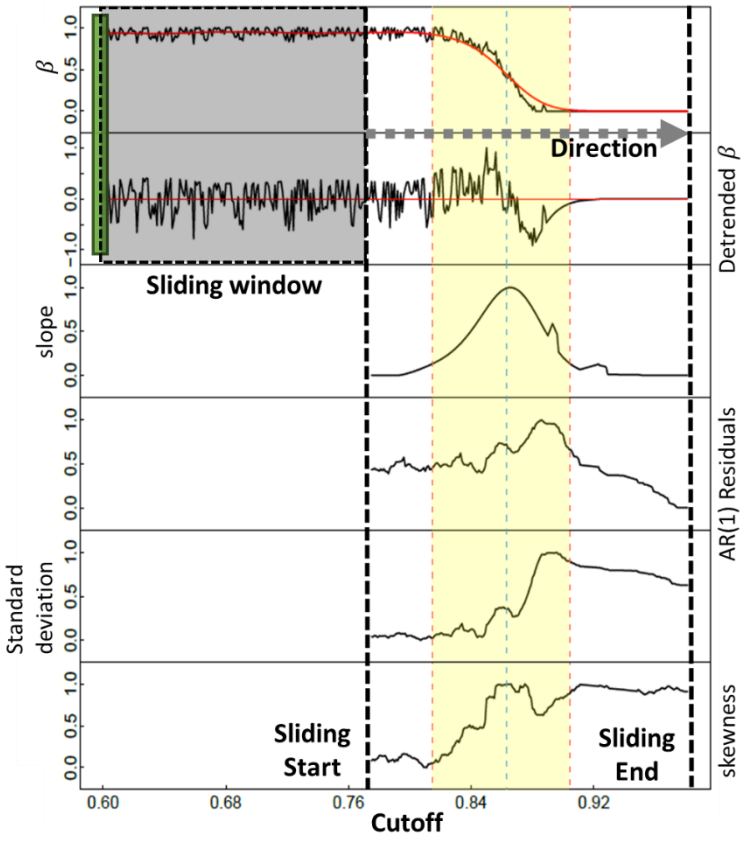


Figure 5.3 An example of critical transition detection and threshold selection using trend analysis on the β dynamics in a numerically simulated system (normal distribution).

5.4.4 Threshold detection comparison with the MENAP

A total of 500 MENAP data association (the PCC) matrix datasets were used for the threshold detection comparison between the GBD-RMT approach and the MENAP. We ran both methods on all of the datasets using the default settings of each, and based on yields from both methods, we evaluated the capability of each method to detect thresholds and compared values of thresholds detected by both methods. As shown in **Figure 5.4**a, both methods have been able to detect the thresholds for a vast majority of the datasets. However, the GBD-RMT approach was able to detect the thresholds in all the datasets successfully, while the MENAP failed to detect the thresholds in a total of 20 (4%) datasets. The thresholds detected by both methods have the similar range of

value, as all of them fell between ~0.5 and 1 (**Figure 5.4**b). Also, the values of thresholds are also similar as indicated by the strong linear correlation (Pearson's *r* =0.85) with a slope close to 1 (**Figure 5.4**b). About 75% of thresholds have differences of value less than 0.025. This result showed that the critical thresholds existed and were detectable to both the RMT approaches for microbial data association networks, and values of detected critical thresholds by both approaches are similar. Second, the GBD-RMT approach has a broader range of detection than the MENAP, because it was able to detect the critical transitions and identify the critical thresholds in the systems where the MENAP failed. Third, the final thresholds selected by the GBD-RMT approach has the most values similar to the ones selected by the MENAP, which suggested that the two different approaches that relied on the same basis of the RMT are consistent in threshold identification.

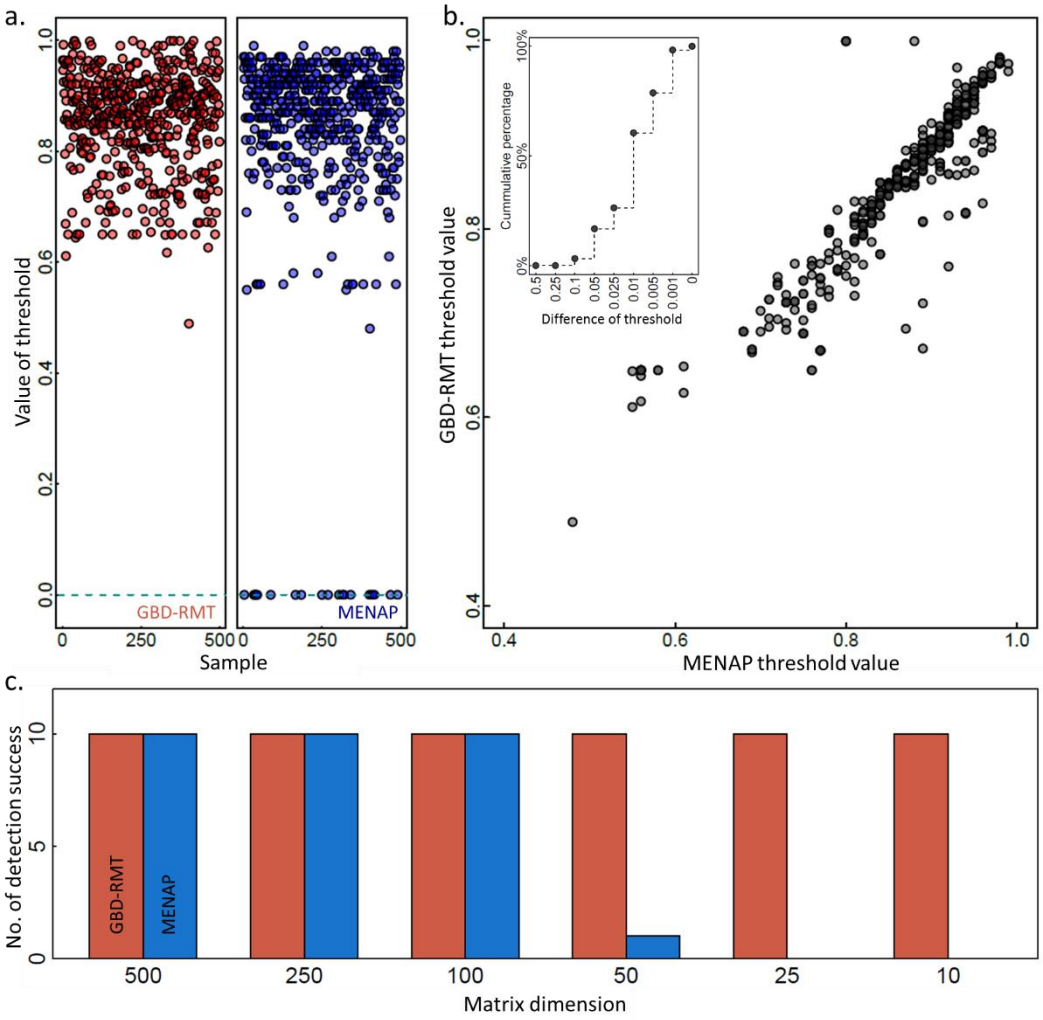


Figure 5.4 (a) Detection of critical threshold for 500 datasets using the GBD-RMT approach and the MENAP. Detection failures were at blue dashed line with zero threshold value. (b) Comparison between values of critical thresholds detected by the GBD-RMT approach and the MENAP. The inner figure shows cumulative percentage of differences between thresholds detected by the GBD-RMT approach and the ones detected by the MENAP. (c) Comparison of the resolution of detection between the GBD-RMT approach and the MENAP in matrices with decreasing dimensions.

The detection of the critical thresholds based on the RMT is dependent on the spectral analyses of the data association matrices, which usually required sufficient non-zero eigenvalues. When the data association matrices have low dimensions, the corresponding non-zero eigenvalues will be less in number, and thus can cause uncertainties in spectral analysis and lead to failure in the detection of the critical

transitions. We ran both the GBD-RMT approach and the MENAP on the data association matrices with different numbers of nodes, to compare the detection limit of the two approaches in terms of the dimension of input matrix. We used the data association (CLR_PCC) matrix (n=2099) calculated based on 16S profiling data from BCI site (the project of biogeographic survey) as a model system. We then generated a total of 60 matrices with 6 lower dimensions (n=500, 250, 100, 50, 25 and 10), that is 10 matrices for each dimension. Each single matrix was generated by randomly drawing the corresponding number of rows and columns from the model system, and thus was different from each other. The results showed that both GBD-RMT approach and the MENAP could detect the critical thresholds in all replicates of 500×500, 250×250 and 100×100 matrices. As the dimension continued to decrease, the GBD-RMT approach and the MENAP started to differ in the resolution and capability of detection. In 50×50 matrices, the MENAP is only capable of detecting the critical thresholds in one matrix, which suggested a loss of resolution of detection of the MENAP. In 10×10 and 25×25 matrices, the MENAP was not capable of yielding any critical thresholds, while the GBD-RMT approach detected thresholds in all matrices in spite of the extremely low dimensions of the matrices. These results suggested that the GBD-RMT approach was less limited in detection of the critical thresholds and higher in resolution of detection in matrices with low dimensions than the MENAP.

5.4.5 Threshold detection comparisons among data association methods

The calculation of data association matrix is an important step in general data association network inference, we also expected that distinctive methods for estimating data association will produce different data association matrices, and consequently

affect the detection of critical thresholds and structures of inferred networks. First, the results (**Figure S 12**) showed that the methods differed in detected association strengths using the abundance data of the model system (BCI). Three methods (dCor, MI, and MIC) don't distinguish the negative associations from the positive ones, and thus had a range of association strength from 0 to 1. The method of LSS detected few values of data association close to 0, due to the nature of this method to select the maximal sum of the product of all possible subsequences. Two methods (MI and MIC) were capable of yielding high strengths (>0.5) of the data associations where other method scoring low (close to 0). The two rank correlation methods, SPM and KDL, detected data association strengths that are highly correlated with each other (Pearson's r =0.99). The CLR_PCC, though derived from the PCC, detected association strengths that were visually different from the PCC, rather more similar to the SPM and KDL. Then we run the GBD-RMT approach on the data association matrices of the same model system calculated using the methods. The results showed that the detected critical thresholds are quite different among the data association methods. The highest critical threshold (0.868) was detected in the MIC dataset, and the lowest one (0.68) in the KDL dataset (**Table 5.1**). Similar conclusions were found statistically significant by comparing critical thresholds detected in all datasets from the project of biogeographic survey and plant succession (**Figure S 13**). More details can be found in **Table S 6**. Next, we constructed the data association networks for each method using the corresponding critical thresholds, and compared the methods by analyzing edge overlap ratio among these networks. The edge overlap ratio here was calculated by dividing the number of edges two methods agreed on both existence and non-existence with the total number of

possible edges. The results of the analysis of the model system indicated that the SPM and KDL networks had the highest edge overlap ratios (>94%) among all pairs, and they are also the two networks that were most similar to the CLR_PCC network (>71%) in terms of common edges. The PCC and dCor networks had edge overlap ratio about 70%, but they are quite different from all other networks (<40% and <35%) in general. It also confirmed that the CLR_PCC network is slight more similar to the SPM and KDL (both >71%) networks than to the PCC network (~40%, **Figure 5.5**). The LSS, MI and MIC networks were highly similar to each other with the edge overlap ratios more than 90% between any pair of the three networks (**Figure 5.5**). We extended the similar analysis to all the datasets from the project of biogeographic survey and plant succession, and found the similar results (**Figure S 14**) in general. Furthermore, we analyzed the differences of the network topological properties of network inferred using different methods for the model system, including the total number of connected nodes, total number of edges, average connectivity, average shortest path, diameter, average clustering coefficient and modularity, and evaluated how these properties differ by methods. The results showed that the networks of different data association methods had also disparate topological properties. The MI network of the model system is the simplest of all, with the lowest number of connected nodes and edges, average connectivity, diameter, and average clustering coefficient. The dCor network of the model system had the highest number of connected nodes and edges, and average connectivity of all, but its modularity is the lowest. Then the similar calculation of the topological properties was made for the networks of all datasets from the project of biogeographic survey and plant succession to statistically examine the differences of

topological properties among all methods. The MI networks had significantly ($p < 0.05$) lower values in five topological properties than almost all of the other methods, except than the LSS in average shortest path and diameter, which is similarly low as the MI. The MIC, MI and LSS had significantly ($p < 0.01$) lower number of connected nodes than the rest of the methods, meanwhile the rest of the methods are not really different from each other. The MIC is significantly ($p < 0.05$) lower than the other methods (except the MI) in the number of edges and average connectivity, but higher than the other methods in modularity (**Figure S 13**).

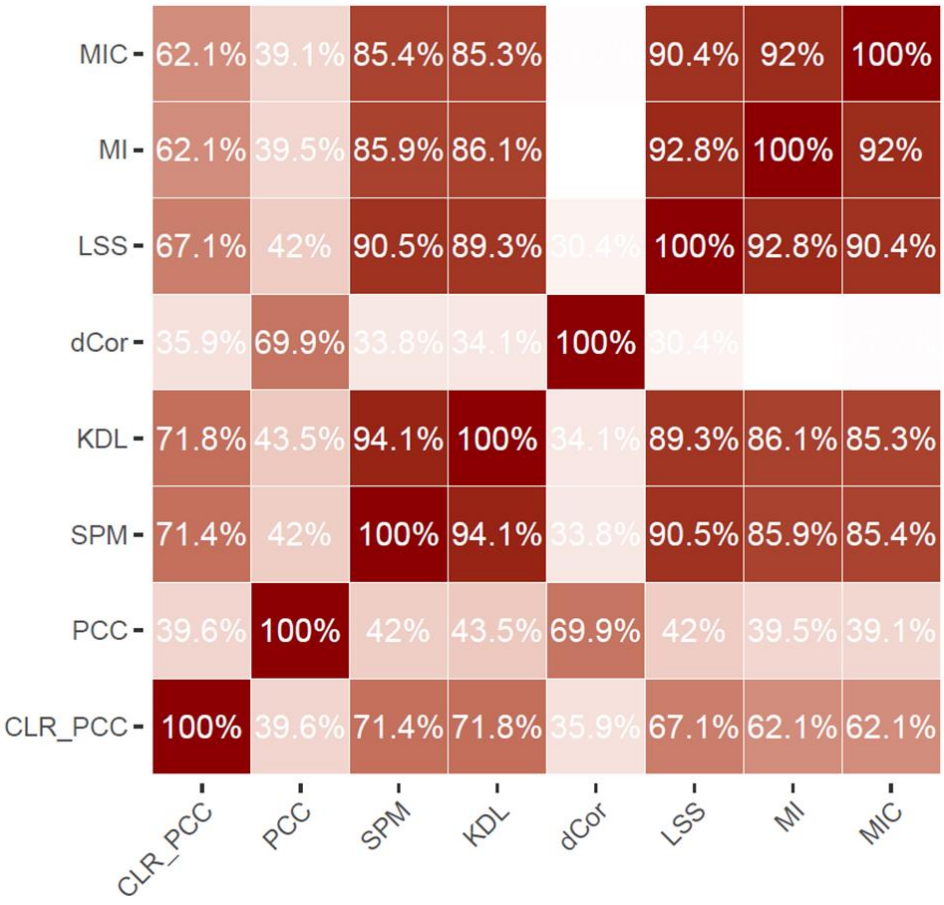


Figure 5.5 Edge overlap ratio among the networks of a model system (BCI) inferred based on different data association detection methods.

Table 5.1 Comparison of detected critical transitions and thresholds, and topological properties among the networks of a model system (BCI) inferred based on the CLR_PCC.

Property	Data association detection method							
	CLR_PCC	PCC	SPM	KDL	dCor	LSS	MI	MIC
Critical transition start	0.672	0.676	0.682	**0.569**	0.68	0.662	**0.732**	0.682
Critical transition end	0.866	0.882	0.847	**0.714**	0.855	0.876	0.816	**0.917**
Critical threshold	0.796	0.788	0.813	**0.68**	0.761	0.86	0.801	**0.868**
No. of connected node	1080	1139	725	783	**1302**	388	**363**	631
No. of edge	5799	9264	1953	1883	**11321**	859	**442**	883
Avg. connectivity	5.58	8.92	1.88	1.81	**10.90**	0.83	**0.43**	0.85
Avg. shortest path	**4.50**	6.01	5.10	5.72	4.84	5.62	5.81	**11.27**
Diameter	**13**	22	14	19	16	16	**13**	**34**
Avg. clustering coeff.	0.34	**0.46**	0.32	0.27	0.43	0.34	**0.08**	0.33
Modularity	0.50	0.51	0.64	0.62	**0.47**	0.66	0.77	**0.80**

5.4.6 Scale-freeness

The scale-freeness is an important topological property widely observed in many complex networks including those from biology. We analyzed the scale-freeness of data association networks generated with varying cutoff values, and the results showed a stunning overlap between the range of cutoff values in the critical transition detected using the GBD-RMT approach and the range of those in which the scale-freeness were high and had low variance (the scale-freeness plateau). The **Figure 5.6**a showed an example of such an observation in our model system (the BCI network based on CLR_PCC). Using different cutoff values were found to greatly affect the scale-freeness (from 0 to 1) of the inferred network (**Figure 5.6**a) The changes of the scale-freeness based on single cutoffs didn't show a clear trend in cutoff values ranged from 0.2 to 0.4, but did show an increasing trend from cutoff of 0.4 to cutoff of 1. Within the critical transition (0.672 to 0.866), the scale-freeness reached high in values (~1) and its curve became visually flat (**Figure 5.6**a). Next, we used a window with a size the same as the spanning of the critical transition to slide on the scale-freeness curve, and then

calculated the mean and variance of the scale-freeness in the window. The results showed an increasing trend of the mean of the scale-freeness (from ~0.3 to ~0.94) and a decreasing trend of the variance (from ~0.18 to 0.02), as the window slid from the beginning of the cutoff values to completely matching the critical transition (**Figure 5.6**b). To confirm the critical transition had significantly higher scale-freeness and lower variance of that, we performed a statistic test in which the entire scale-freeness series were permutated 9999 times, and each time, the mean and variance of the permutated scale-freeness in the critical transition were calculated and compared with the real ones. As results, the permutation test showed that the scale-freeness in the critical transition had significantly higher mean value ($p < 0.0001$) and lower variance ($p < 0.0001$) (**Figure 5.6**c and d). The above analyses have been performed on all datasets from the project of biogeographic survey and the project of plant succession, and similar result were observed for each individual dataset (**Table 5.2**), which implied that the overlap between critical transitions and the range of those in the scale-freeness plateau existed pervasively in the analyzed biological networks despite different microbial communities and data association detection methods.

Table 5.2 Overlaps the critical transitions and the scale-freeness plateaus in all datasets based on the CLR_PCC.

Sample ID	Span of critical transition	Average scale-freeness in critical transition	Variance of scale-freeness in critical transition	Scale-freeness at critical threshold	Average scale-freeness of all	Variance of scale-freeness of all	Largest scale-freeness of all	Is largest scale-freeness in critical transition	p value of null model for mean and variance of scale-freeness	
									mean	variance
AND	0.186	0.964	0.018	0.947	0.704	0.282	0.996	Yes	**<0.0001**	**<0.0001**
BCI	0.246	0.955	0.029	0.961	0.726	0.264	1.000	Yes	**<0.0001**	**<0.0001**
CWT	0.235	0.960	0.035	0.953	0.731	0.245	1.000	Yes	**<0.0001**	**<0.0001**
HFR	0.194	0.939	0.028	0.973	0.646	0.295	0.993	Yes	**<0.0001**	**<0.0001**
LUQ	0.196	0.960	0.024	0.986	0.618	0.332	0.997	Yes	**<0.0001**	**<0.0001**
NWT	0.174	0.971	0.023	0.987	0.662	0.298	0.997	Yes	**<0.0001**	**<0.0001**
S1W0	0.088	0.915	0.061	0.907	0.426	0.297	1.000	No	**<0.0001**	**<0.0001**
S1W3B	0.086	0.982	0.014	0.990	0.650	0.253	1.000	Yes	**<0.0001**	**<0.0001**
S1W6B	0.16	0.954	0.017	0.953	0.687	0.326	0.999	No	**<0.0001**	**<0.0001**
S1W9B	0.078	0.863	0.093	0.927	0.358	0.288	0.998	No	**<0.0001**	**<0.0001**
S1W12B	0.093	0.959	0.020	0.959	0.532	0.289	0.998	Yes	**<0.0001**	**<0.0001**
S1W3R	0.077	0.914	0.071	0.978	0.560	0.232	1.000	No	**<0.0001**	**<0.0001**
S1W6R	0.132	0.948	0.020	0.945	0.703	0.289	1.000	No	**<0.0001**	**<0.0001**
S1W9R	0.085	0.973	0.018	0.983	0.616	0.289	0.997	Yes	**<0.0001**	**<0.0001**
S1W12R	0.121	0.966	0.022	0.984	0.766	0.223	1.000	No	**<0.0001**	**<0.0001**
S2W0	0.099	0.977	0.014	0.981	0.593	0.334	0.999	No	**<0.0001**	**<0.0001**
S2W3B	0.08	0.951	0.026	0.978	0.475	0.311	1.000	No	**<0.0001**	**<0.0001**
S2W6B	0.132	0.953	0.034	0.970	0.734	0.251	1.000	No	**<0.0001**	**<0.0001**
S2W9B	0.111	0.937	0.032	0.953	0.747	0.246	0.997	No	**<0.0001**	**<0.0001**
S2W12B	0.088	0.981	0.015	0.991	0.571	0.296	0.999	Yes	**<0.0001**	**<0.0001**
S2W3R	0.102	0.972	0.024	0.996	0.566	0.278	1.000	Yes	**<0.0001**	**<0.0001**
S2W6R	0.086	0.985	0.012	0.987	0.516	0.338	1.000	Yes	**<0.0001**	**<0.0001**
S2W9R	0.089	0.974	0.029	0.991	0.643	0.241	1.000	Yes	**<0.0001**	**<0.0001**
S2W12R	0.134	0.979	0.017	0.997	0.746	0.228	0.999	Yes	**<0.0001**	**<0.0001**

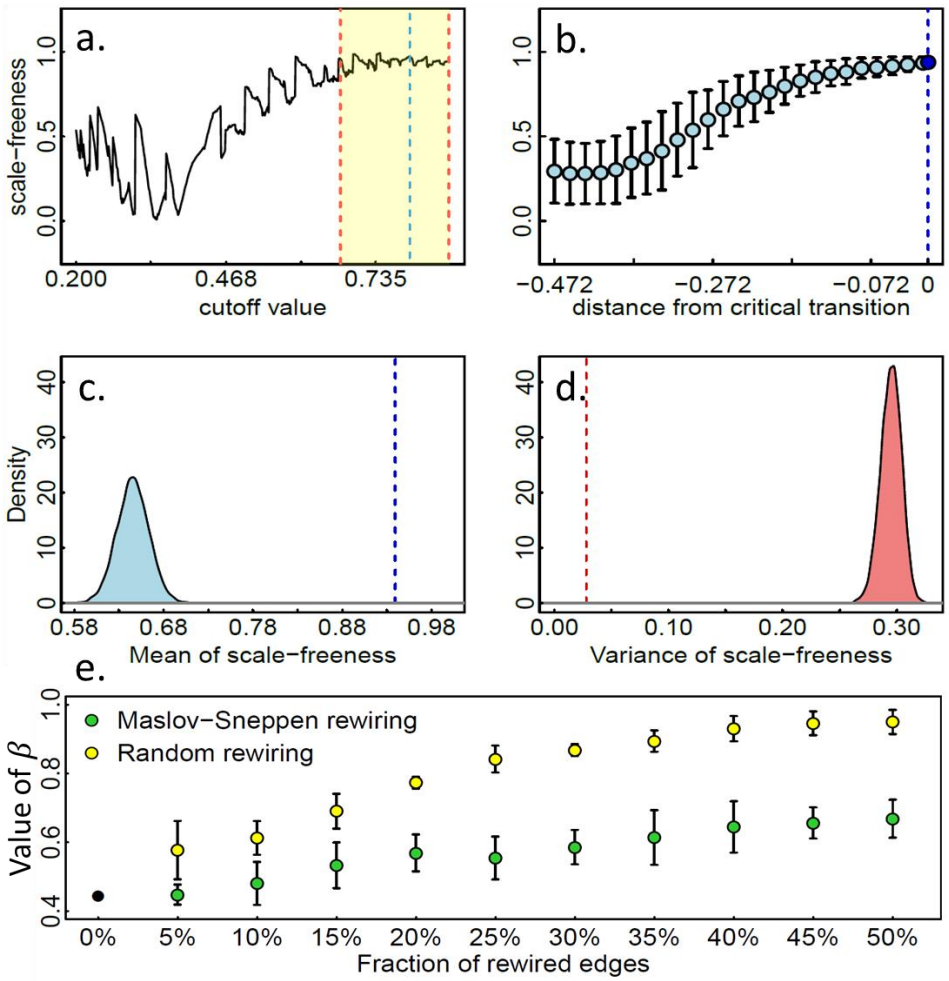


Figure 5.6 An example of the overlap between the critical transition and the scale-freeness plateau in a model system (BCI) based on the CLR_PCC method. (a) Changes of scale-freeness with increasing cutoff values. The critical transition region is in yellow which is in between two red dashed lines. The critical threshold is indicated by the blue dashed line. (b) Mean and variance of scale-freeness in a window with the same spanning as the critical transition, whose left side slides from beginning cutoff (0.2) to the beginning of the critical transition (0.672). The blue dashed line indicates those in the critical transition. (c) Permutation (n=9999) test to verify mean of scale-freeness in the critical transition is significantly higher than what from other regions. Means of scale-freeness from permutations are distributed in light blue shape, the original mean of scale-freeness is indicated by the blue dashed line. (b) Permutation (n=9999) test to verify variance of scale-freeness in the critical transition is significantly lower than what from other regions. Variances of scale-freeness from permutations are distributed in light red shape, the original variance of scale-freeness is indicated by the red dashed line. (e) Changes of β value of the model system at the critical threshold in response to the fractions of rewired edges with two different rewiring procedures. Both the procedures randomly change the organization of and preserves the number of edges.

Furthermore, we analyzed how the β value obtained with the critical threshold responded to the changes of the scale-freeness. In the analysis, we used two different procedures, the Maslov-Sneppen rewiring procedure [56] and the random rewiring procedure, to rewire a gradient of percentage (5%, 10%, 15%, 20%, 25%, 30%, 35%, 40% and 50%) of edges in the data association network constructed using the critical threshold. In the Maslov-Sneppen procedure, two edges without shared nodes were selected and rewired at the same time, and the rewiring will only change one each time. Both of the procedures preserve the number of edges, but the Maslov-Sneppen procedure keeps the degree distribution of the network unchanged and the random rewiring does not. So the scale-freeness of a network will only change when the random rewiring is applied, and it will not when using the Maslov-Sneppen procedure. For each rewired network, we reversely obtained the corresponding data association matrix, and then estimated the corresponding β value using the GBD-RMT approach to see how it changes in comparison with the original β value (~0.44) obtained with the critical threshold (0.796). The results first showed a trend that values of β parameter increased as more edges were rewired in both the Maslov-Sneppen procedure and the random rewiring procedure, which suggested that both the rewiring procedures caused the NNSD of the system to shift towards the Wigner-Dyson distribution and thus an increase of randomness in the system. Interestingly, the result of two procedures differed in how much values of β parameter were changed when edges were rewired. The value of β, in the random rewiring procedure, was increased to ~0.7 (57% increase) when a small fraction (15%) of edges were rewired, and increased to ~0.9 (103% increase) for 35% rewired edges. In the Maslov-Sneppen procedure, the value of β

increased to 0.67 (52% increase) where the highest fraction of edges was rewired in our test, which was less than the increase by rewiring 15% of edges in the random rewiring procedure. These results suggested that the β value obtained with the critical threshold is sensitive to the changes of the scale-freeness.

5.5 Discussion

5.5.1 Importance of network analysis

Characterizing complex microbial interactomes is critical to understanding microbial diversity and function (Zhou, Wu et al. 2011, Deng, Jiang et al. 2012, Shi, Nuccio et al. 2016), but most previous microbial ecology studies only focused on simple microbial richness and composition. Ignoring the organization and dynamics of microbial interactions makes it hardly possible to fully understand diversity and species co-evolution in microbial communities. In recent years, network based analysis has become an emerging tool for exploring microbial interactions in complex microbial communities. Previously, we systematically described a data association network approach based on the mathematical framework of the RMT, demonstrated its applicability in characterizing interactomes in different biological systems (e.g. protein interaction network (Luo, Yang et al. 2007), functional gene network (Zhou, Deng et al. 2010) and phylogenetic network (Zhou, Wu et al. 2011)), and also presented a bioinformatics tool, the MENAP (Deng, Jiang et al. 2012). While the MENAP has been powerful in objectively selecting critical thresholds for inferring correlation networks in microbial communities, it still has limitations in dealing with compositional data, detecting non-PCC associations, and identifying and interpreting of critical thresholds. In this study, we presented the GBD-RMT approach, which was based on the same

RMT framework but with new methods and improvements for better inferring data association networks.

5.5.2 Advantages of this approach

As an approach based on the RMT framework, the GBD-RMT approach detected threshold values that are similar to the ones selected by the MENAP, which suggested that the two approaches that relied on the same framework are consistent with each other. Both the approaches share the advantages of objective detection of critical thresholds for inferring data association networks. Comparing with the MENAP, the GBD-RMT approach had several advantages.

First, the GBD-RMT approach is better than the MENAP in the capability and resolution of critical threshold detection. The MENAP has been able to detect critical thresholds in the most datasets, but it still failed a few. While the failure rate is not high, it is likely to preclude comparative analysis on multiple networks, especially when there were many networks to compare at the same time. In such cases, one single detection failure causes imbalanced comparisons, and thus affects further analysis and conclusion. The detection failure is likely caused by two reasons. The MENAP approximated the NNSD by binning discrete spacing data points, whose performance was likely affected by the number of bins, thus it can cause inaccurate approximation when the sub-optimal bin number was used and further affect the characterization of NNSD. Meanwhile, the MENAP makes threshold calling based on the *p* value of fitting the NNSD to Wigner-Dyson and Poisson distribution in χ^2 test, which is highly sensitive to the number of data points and small frequencies (Fornell and Larcker 1981) in the NNSD. When candidate cutoff value is high, the non-zero eigenvalues of the reduced data association

matrix will be less, and thus the NNSD will have fewer data points and be more likely to have small frequencies, which lead to erroneous conclusions in threshold calling. The presented GBD-RMT approach overcomes these limitations by use MLE based estimation parameter on the NNSD, which avoids both the NNSD binning and χ^2 test and have additional advantages in consistency, efficiency, and invariant (Myung 2003). Consequently, it succeeded in detecting critical thresholds in all simulated and real datasets throughout this study, including those where the MENAP failed, and also had higher detection resolution than the MENAP. The perfect detection by the GBD-RMT approach not only once again confirmed the applicability of the RMT based approach in biological systems, but also provided an important basis for a large scale comparative network analysis, which is anticipated in metagenomics enabled studies as cost of microbial community profiling keeps on being brought down by advances in the metagenomics technology. In the meantime, the advantage of higher detection resolution makes the GBD-RMT approach more appropriate for inferring network from biological systems (e.g. less diverse microbial communities) that are less complex in terms of having fewer nodes.

Second, detection of critical threshold in the GBD-RMT approach has higher confidence and interpretability than what in the MENAP. Apart from the MENAP, which calls threshold on each candidate cutoff in isolation from all other cutoffs, the GBD-RMT approach recovers entire β dynamics first, and then analyze the trend of the β dynamics for detecting critical transitions and selecting critical threshold. Since each value in the β dynamics is estimated from the GBD, which quantitatively measures where the NNSD stands in between the Wigner-Dyson distribution and the Poisson

distribution, the GBD-RMT approach can detect critical transition based on transition progress, where the trend of β dynamics starts to decrease from 1 but before dropping to 0. In the RMT context, systems experiencing critical transition have the randomness that is no longer separable from the system specific properties, so all cutoffs fallen in the critical transition have the potential to be the critical threshold. In this study, the cutoff value at which the slope of β dynamics trends was the highest was empirically selected as the critical threshold, which means critical transition occurs at the fastest pace at the critical threshold. In practice, selections for the critical threshold with good interpretability includes beginning and ending cutoff of critical transition, and cutoff at the middle of overlap of critical transitions. The principle is that all thresholds should be selected in the same way in terms of interpretation, otherwise multiple networks might be compared on different basis and thus the conclusion of the analysis can be affected. In contrast, the MENAP calls thresholds when target NNSDs fit the Poisson distribution without giving any information about the progress of transitions of NNSDs and without knowing states of NNSDs before the thresholds to call and thereafter. In this manner, even though the critical thresholds identified by the MENAP are most likely from the critical transitions, it is uncertain or less confident than the GBD-RMT approach because the exact states of system NNSDs are not known. We showed the critical transitions can have long spans, so when comparing multiple networks, ignoring what states of NNSDs for which the detected thresholds stand is likely to lead networks to compare to be inferred with criteria that are not unified, and thus cause not only uncertainty and but also interpretability issue in comparisons.

In addition, the GBD-RMT approach integrated the state-of-art techniques for tackling several challenges faced with the inference of data association networks, which were lacked in the MENAP. The GBD-RMT approach implemented with the clr, which alleviate the compositional data bias for inferring correlation networks, while the MENAP doesn't. This is important, especially to the network analysis in microbial ecology studies, because absolute abundance data is difficult to access and compositional data (e.g. microbial profiling data based on high-throughput sequencing) is prevalent in the field. Therefore, incorporating the clr should improve the accuracy of the correlation network inference, and make the GBD-RMT approach more prepared than the MENAP for in those studies. Meanwhile, the GBD-RMT approach integrated multiple methods for detecting the data association between two random variables, while the MENAP is solely based on the PCC. Each method has distinctive features, and thus produces data association matrices with different critical thresholds and infers networks with different structures and topological properties in our results. The distinctiveness of the SPM (Barberán, Bates et al. 2012, Szklarczyk, Franceschini et al. 2014), KDL (Das, Meher et al. 2017), dCor (Guo, Zhang et al. 2014), LSS (Steele, Countway et al. 2011), MI (Song, Langfelder et al. 2012) and MIC (Reshef, Reshef et al. 2011) demonstratively makes each of them work well or perform better than others in inferring networks from specific complex biological systems. The variety of data association methods gains a practical advantage for the GBD-RMT approach in inferring networks from a broad range of datasets. For the systems where only linear data associations are important, that is other types of association are considered as noise, methods like the PCC should be chosen over the LSS, MI and MIC, which is

highly and only sensitive to linear dependences (Speed 2011); on the contrary, if as many types of associations as possible are interested, the MI and MIC should be first considered, which reportedly outperformed others in detect non-linear or non-functional associations (Reshef, Reshef et al. 2011, Kinney and Atwal 2014). Meanwhile, if both negative and positive data associations are interested, the dCor, MI and MIC may be inappropriate as they don't assign signs to the detected data association edges as shown in our results.

5.5.3 Scale-freeness

The scale-freeness is an important topological property, which differentiates many complex networks from randomly generated networks (Barabási and Albert 1999). The scale-freeness has received a lot of attention because many notable characteristics (e.g. robustness to failure) are commonly shared by the networks that are scale-free. Whether a network is scale-free or not alone help us better understand the formation of the network and the factors that shape network structure, and predict influences of network on how system functions respond to disturbances (Albert, Jeong et al. 2000, Pastor-Satorras and Vespignani 2001, Eguiluz, Chialvo et al. 2005). In biological systems, the scale-free networks also widely exist in interactomes of microbial species, metabolites, functional genes, and proteins (Jeong, Mason et al. 2001, Ravasz, Somera et al. 2002, Barabási and Oltvai 2004, Albert 2005, Chaffron, Rehrauer et al. 2010, Zhou, Deng et al. 2010). With the GBD-RMT approach, we found, for the first time, a remarkable linkage between the critical transition of β series and the plateau of scale-freeness, which pervasively existed in the biological systems included in this study. The overlap found in this study has several important implications. First, the scale-freeness depends

on the selection of the critical threshold. Therefore, selecting a critical threshold arbitrarily can lead to very different conclusion about the scale-freeness of the same network, which also emphasized the importance of the methods detecting critical thresholds objectively. Second, the high scale-freeness uncovered by the GBD-RMT approach for all the datasets is consistent with the prevalence of scale-free networks in biological systems. This scale-free nature of complex biological networks can be explained by the preferential attachment hypothesis, which generalizes mechanisms that drive new nodes to preferentially connect to the existed nodes with high degree during network formation or growth. Here, the mechanism we think explains the scale-freeness of microbial networks is functional redundancies. In a microbial community, multiple species can contribute equivalently to an ecological function, so these species that have duplicate functions are likely to interact with the same partners. Intuitively, labile Carbon (C) decomposers of difference species, for example, are likely to interact with upstream recalcitrant C decomposers, so each recalcitrant C decomposer gains a new link to all species that decomposes labile C. In this way, the scale-free network is like to form if there exist important but less redundant functions in the microbial community. Third, the linkage itself is consistent enough in our analysis to drive us to hypothesize its universality among all biological or non-biological data association networks. To further test the hypothesis, it requires a larger scale analysis on microbial data association networks or rigorous mathematic proof. If the hypothesis holds, the linkage should be a very interesting connection between the RMT and Network Science.

5.5.4 Limitations and future work

The GBD-RMT approach is free of limitations. First, as an approach for inferring the networks based on data association, the GBD-RMT approach inherently has the common limitations. The edges among the nodes in the inferred networks should be only considered as the hypothetical microbial interactions, which by no means automatically extends to any empirical validation on the microbial interactome, nor implicates any causality or direction information. Second, the GBD-RMT approach is not able to solve compositional data bias for data association methods other than the PCC. So, the networks inferred with those methods will be inevitably biased, and future effort is still needed for determining how the compositional data bias affects each individual data association method and developing corresponding bias reduction methods. Third, the GBD-RMT approach is not capable of decisively determining which particular data association method should be preferred over others to a task of network inference, without additional information about what types of data association are of importance and interest in the particular system may be highly necessary to determine so. To further lift the limitation, it requires experimental verifications to establish linkages between features of abundance data and selection of data association method. Finally, the GBD-RMT approach is limited in speed comparing with the MENAP. Both the spectral analysis and MLE based estimation for an entire β dynamics in the GBD-RMT approach are computationally intensive. The time cost of both procedures was easy to see to depend on the scale of cutoff ranges. The speed of the spectral analysis was also further limited by the dimension of th4e data association matrix, because the computational complexity of solving an n-dimension matrix with

low sparsity for eigenvalues is O (n^3). To get more favorable speed, the approach was implemented in parallel computing scheme, which allows performing the procedures for multiple cutoffs concurrently and brings down the time cost no more than the MENAP. But it still needs further speed-up for processing large data association matrices (n > 10,000). Possible solutions may include predicting a narrower range of candidate cutoffs before detecting the critical transition based on training datasets, which will be explored in our future work. Despite all these limitations, the GBD-RMT approach is still powerful in providing insights into the microbial interactomes, generating hypotheses for further testing, and adding a substantial dimension to microbial ecology studies beyond those of simple analysis of richness and composition.

Chapter 6: Summary and Output

This work included the development of two high-throughput FGAs for the specific and general purpose, respectively, a bioinformatics pipeline for function-oriented analysis of shotgun metagenome sequencing data, and a computational approach for inferring data association networks. Each developed technology or method was demonstrated to be powerful in application studies. Overall, the work in this dissertation offered new and up-to-date technological and computational resources for improving our understanding of complex microbial communities.

First of all, we developed the PABMC, since detecting and profiling soil microbial functional genes beneficial to plants is critical to understand interactions between PGPMs and plants in agricultural and natural ecosystems, which is of great economic and ecological importance. To our knowledge, the PABMC is the first high-throughput functional gene array to characterize plant beneficial genes with comprehensive coverage in terms of plant beneficial genes and PGPM species. The specificity of the probes included in the PABMC was verified to be highly specific in the computational evaluation. In the showcase study to investigate PGPM communities in natural sites where have been invaded by *A. adenophora*, The PABMC uncovered the impacts of plant invasion on microbial communities from a perspective of plant beneficial genes, and offered evidence for explaining how invasive plant interacted with and receive benefits from PGPMs and eventually established successional success. The evidence includes the increased alpha diversity and the shifted composition of communities in the invaded site, as well as the increased abundance of a drought tolerance gene and the decreased abundance of pathogen resistance genes. Interestingly, different directions of

significant changes were observed in response to the *A. adenophora* invasion, for antibiotic biosynthesis, in which abundances increased in two genes and decreased in three. We speculated that *A. adenophora* might actively overturn belowground antibiotic regulatory landscape by suppressing original antibiotic synthesizers established in the native site in favor of substitute synthesizers. While further studies are required to examine *A. adenophora* root exudates, soil properties and the surrounding microbiota in order to confirm these hypotheses, the PABMC was demonstrated to be a powerful tool for efficient characterization of PGPM.

Second, to further address various experimental and computational challenges still existed in analyzing microbial communities in the environment, we have developed a new generation of functional gene arrays (GeoChip 5.0). The GeoChip 5.0 contains 161,961 probes covering functional groups involved in microbial carbon (C), nitrogen (N), sulfur (S), and phosphorus (P) cycling, energy metabolism, antibiotic resistance, metal resistance/reduction, organic contaminant remediation, stress responses, pathogenesis and virulence as well as markers specific for viruses, protists, and fungi. To the best of our knowledge, this is the most comprehensive functional gene arrays currently available for studying microbial communities important to biogeochemistry, ecology, environmental sciences as well as human health. Compared with previous generations of GeoChip, GeoChip 5.0 has several improved features, such as covering novel functional categories (e.g. microbial defense, protist, plant growth promotion, pigments and metabolic pathways), targeting more functional gene families and genes, and having smaller spots with higher density. Meanwhile, both computational and experimental evaluations demonstrated that the developed Agilent-based GeoChip 5.0 is

highly specific, sensitive, and quantitative for characterizing microbial community functional composition and structure. Furthermore, in our application study, GeoChip 5.0 was used to examine the impacts of heavy metal contamination on groundwater microbial communities. The result uncovered decreased functional gene richness and Shannon-Wiener diversity, shifted microbial community composition, and changes in abundance of metal homeostasis genes in the contaminated samples. The application also revealed that environmental variables including U and DOC played critical roles in shaping microbial community structure. All these results demonstrated the capability of GeoChip 5.0 for linking microbial communities to various ecosystem functional processes.

Third, we developed the EcoFun-MAP, with database construction, workflow design and pipeline implementation and deployment. The developed EcoFun-MAP has several unique features and advantages, and thus is capable of addressing some of the computational barriers brought by rapid throughput increase in NGS technology and faced by many microbial ecologists. The databases of EcoFun-MAP are smaller and less in redundancy than general public databases, with a selective coverage and specialized organization of functional genes that are important to microbial ecology studies, and offered in a variety of widely accepted data structures. The quality of reference sequences was ensured using two separate procedures with manual corrections, thus the reference sequences used for database construction should be accurate. Meanwhile, the workflows of the EcoFun-MAP offered exceptional speed, which is, to our best knowledge, the first web-based pipeline with speed of multi-million reads per minute. The EcoFun-MAP gained speed advantages through smaller

reference databases, fast and updated tools and HPC implementation. In addition, all EcoFun-MAP workflows were capable of generating similar results in analyses of composition and functional gene richness of the metagenomes from underground water samples in our analysis. Detailed analyses of the metagenomes based on the EcoFun-MAP demonstrated its usefulness in revealing differences in relative abundances of functional categories and functional genes among sampled microbial communities, and link the differences with different levels of contaminants in the samples.

Fourth, we developed the GBD-RMT approach, which was based on a previous RMT framework but with new methods and improvements for better inferring data association networks. The GBD-RMT approach detected threshold values that are similar to the ones selected by the previous RMT-based approach, the MENAP, which suggested that the two approaches that relied on the same framework are consistent with each other. However, comparing with the MENAP, the GBD-RMT approach had several advantages. The GBD-RMT approach is better than the MENAP in the capability and resolution of critical threshold detection, because it used MLE based estimation parameter on the NNSD, which avoids both the NNSD binning and χ^2 test and have additional advantages in consistency, efficiency, and invariant. Meanwhile, detection of critical threshold in the GBD-RMT approach has higher confidence and interpretability than what in the MENAP. Apart from the MENAP, which calls threshold on each candidate cutoff in isolation from all other cutoffs, the GBD-RMT approach recovers entire β dynamics first, and then analyze the trend of the β dynamics for detecting critical transitions and selecting critical threshold. In addition, the GBD-RMT approach integrated the state-of-art techniques for tackling several challenges

faced with the inference of data association networks, which were lacked in the MENAP, including compositional data bias and complex data association types. With the GBD-RMT approach, we found, for the first time, a remarkable linkage between the critical transition of β series and the plateau of scale-freeness, which pervasively existed in the biological systems included in this study. Despite the limitations, such as inferring no directions and causality, the GBD-RMT approach is still powerful in providing insights into the microbial interactomes, generating hypotheses for further testing, and adding a substantial dimension to microbial ecology studies beyond those of simple analysis of richness and composition.

In conclusion, this work provided powerful, novel and update-to-date high-throughput metagenomics technologies, bioinformatics tools and computational methods for analyzing complex microbial communities and interactomes, which adds important parts into the integrated omics toolbox for microbial community analyses

Listed below are the individual studies which were in relation to this work:

1. Shi, S., Nuccio, E., **Shi, Z.**, He, Z., Zhou, J., and Firestone, M. (2016), "The interconnected rhizosphere: High network complexity dominates rhizosphere assemblages", *Ecology Letters*.
2. Xue, K., Yuan, M., **Shi, Z.**, Wu, L., He, Z., Bracho, R., Natali, S., Schuur, E., Luo, C., Konstantinidis, Wang, Q., K., Cole, J., Tiedje, J., Luo, Y., and Zhou, J. (2016), "Tundra soil carbon is vulnerable to rapid microbial decomposition under climate warming", *Nature Climate Change*.
3. Tu, Q., Li J., **Shi, Z.**, Chen Y., Lin, L., Li J., Wang H., Yan J., Zhou Q., Li X., Li L., Zhou J., He Z. (2016), "*HuMiChip2* for strain level identification and functional profiling of human microbiomes", *Applied Microbiology and Biotechnology*.
4. Zhang, Y., Yang Q., Ling J., Van Nostrand, J., **Shi, Z.**, Zhou, J., Dong J. (2016), "The shifts of diazotrophic communities in spring and summer associated with coral Galaxea astreata, Pavona decussata and Porites lutea", *Frontiers in Microbiology.*
5. Hemme, C., Tu, Q., **Shi, Z.**, Qin, Y., Gao W., Deng Y., Van Nostrand, J., Wu, L., He, Z., Chain, P., Tringe, S., Fields, M., Rubin, E., Tiedje, J., Hazen, T., Arkin, A. and Zhou, J. (2015), "Comparative metagenomics reveals impact of contaminants on groundwater microbial communities", *Frontiers in Microbiology*.

6. Zhang, Y., Ling, J., Yang, Q., Wen, C., Yan, Q., Sun, H., Van Nostrand, J., **Shi, Z.**, Zhou, J. and Dong, J. (2015), "The functional gene composition and metabolic potential of coral-associated microbial communities", *Scientific Report*.
7. Zhang, Y., Tian, Z., Liu, M., **Shi, Z.**, Hale, L., Zhou, J. and Yang, M. (2015), "High concentrations of the antibiotic spiramycin in wastewater lead to high abundance of ammonia-oxidizing archaea in nitrifying populations", *Environmental Science & Technology*.
8. Xu, T., Li, Y., **Shi, Z**., Hemme, C., Li, Y., Zhu, Y., Van Nostrand, J., He, X. and Zhou, J. (2015), "Efficient genome editing in *Clostridium cellulolyticum* via CRISPR-Cas9 nickase", *Applied Environmental Microbiology*.
9. Yan, Q., Bi, Y., Deng, Y., He, Z., Wu, L., **Shi, Z**., Li, J., Wang, X., Hu, Z., Yu, Y. and Zhou, J. (2015), "Impacts of the world's largest dam (Three-Gorges Dam) on microbial functions as revealed by comparative Metagenomics". *Scientific Report*.
10. Wang, C., Wang, X., Liu, D., Wu, H., Lü, X., Fang, Y., Cheng, W., Luo, W., Jiang, P., **Shi, Z.**, Yin, H., Zhou, J., Han X. and Bai E. (2014), "Aridity threshold in controlling ecosystem nitrogen cycling in arid and semi-arid grasslands", *Nature Communication*.
11. Luo, C., Rodriguez-R, L., Johnston, E., Wu, L., Cheng, L., Xue, K., Tu, Q., Deng, Y., He, Z., **Shi, J.**, Yuan, M., Sherry, R., Li, D., Luo, Y., Schuur, E., Chain, P., Tiedje, J., Zhou, J. and Konstantinidis K. (2014), "Soil microbial community responses to a decade of warming as revealed by comparative metagenomics", *Applied Environmental Microbiology*.
12. Tu, Q., Yu, H., He, Z., Deng, Y., Wu, L., Van Nostrand, J. D., Zhou, A., Voordeckers, J., Lee, Y.-J., Qin, Y., Hemme, C. L., **Shi, Z.**, Xue, K., Yuan, T., Wang, A. and Zhou, J. (2014), "*GeoChip 4*: a functional gene-array-based high-throughput environmental technology for microbial community analysis", *Molecular Ecology Resources*.
13. Zhou, J., Jiang, Y.-H., Deng, Y., **Shi, Z.**, Zhou, B. Y., Xue, K., Wu, L., He, Z. and Yang, Y. (2013), "Random sampling process leads to overestimation of β-Diversity of microbial communities", *mBio*.
14. He, Z., Zhang, P., Wu, L., Rocha, A., Tu, Q., **Shi, Z**., Qin, Y., Wang, J., Curtis, D., Wu, B., Van Nostrand, J., Wu, L., Yang, Y., Elias, D., Watson, D., Adams, M., Fields, M., Alm, E., Hazen, T., Adams, P., Arkin, A. and Zhou, J., "Microbial functional genes predict groundwater contamination and ecosystem functioning", *in preparation*.
15. Wu L., Chen S., **Shi Z**., Zhao M., Zhu Z., Yang Y., Qu Y., Ma Q., He Z., Zhou J. and He Q., "Microbial functional trait of rRNA operon copy numbers increases with organic levels in anaerobic digesters", *in preparation*.
16. Zhou A., Lau R., Baran R., Ma J., von Netzer F., Shi W., Gorman-Lewis D., He Z., Qin Y., **Shi Z.**, Kempher M., Zane G., Wu L., Bowen B., Northen T., Hillesland K., Stahl D., Wall J., and Arkin A., "Persistent accumulation of key cellular components but altered physiological and transcriptional responses in salt-evolved Desulfovibrio vulgaris", *in preparation*.
17. Chen C., Hemme C., Beleno J., Qin Y., Ning D., **Shi Z.**, Tu Q., Jorgensen M., He Z., Wu L., and Zhou. J., "Disentangling ecological processes underlying formation and changes of microbiomes of periodontal health and chronic periodontitis in response to initial therapy", *in preparation*.
18. Zhang Y., Yang Q., Ling J., Van Nostrand, J., **Shi Z.**, Zhou J., Dong J., "Diversity and Structure of Bacterial Diazotrophic Communities in the Mangrove Rhizosphere as Revealed by Next-generation Sequencing", *in preparation*.
19. Yang Y., Zhang J., Gao Q., Zhang Q., Wang T., Yue H., Wu L., **Shi J**, Qin Z., Zhou J., Zuo J., "Bacteriophage - prokaryote dynamics and interaction within anaerobic digestion processes across time and space", *in preparation*.

20. Gao Y., Ding J., Gu B., Yuan M., **Shi Z.**, Chiariello N., Niboyet A., Le Roux X., Docherty K., Gutknecht J., Hungate B., Yuan T., Gu Y., Field C., Zhou J., and Yang Y., "The effects of long-term warming on soil microbial taxonomic and functional traits in a Mediterranean-type grassland", *in preparation*.
21. Feng J., Yang Y., Penton C., He Z., Van Nostrand J., Yuan M., Wu L., Qin Y., **Shi Z.**, Schuur E., Bracho R., Tiedje J., Konstantinidis K., and Zhou J., "Warming altered nifH-harboring Bacterial Community Composition in Alaskan soils", *in preparation*.
22. **Shi, Z.**, Lu, X., Xue, K., Deng, Y., Van Nostrand, J., Wu, L., Yuan, T., He, Z. and Zhou, J., "Development of a functional gene array for characterizing plant beneficial microorganisms", *in preparation*.
23. **Shi, Z.**, Yin, H., Voordeckers, J., Lee, Y., Deng, Y., Zhou, A., Van Nostrand, J., Wu, L., He, Z., Schuur, E. and Zhou, J., "A specific, sensitive, quantitative and reproducible functional gene array for functionally profiling microbial communities", *in preparation*.
24. **Shi, Z.**, He, Z., Tu, Q. and Zhou, J., "*EcoFun-MAP*: An *Ecological Function Oriented Metagenomic Analysis Pipeline* for agile annotation of deep shotgun sequencing data", *in preparation*.
25. **Shi, Z.**, Xu, T. and Zhou, J., "A web based *Genome Editing Site Analysis System* (*GESAS*) for facilitating prokaryotic genome edition", *in preparation*.
26. **Shi, Z.**, Qin, Y., Xu, Y. and Zhou, J., "A random matrix system approach based on General Brody Distribution for inferring the accurate microbial co-occurrence network", *in preparation*.

Appendix A: Supplementary Figures

Figure S 1 for Chapter 2: Development of a Functional Gene Array to Characterize Plant Growth Promoting Microorganisms Beneficial to Plants

Figure S 2 to **Figure S 6** for Chapter 3: Ultra-sensitive and -quantitative Detection of Microbial Populations in complex communities with New Functional Gene Arrays

Figure S 7 to **Figure S 9** for Chapter 4: The EcoFun-MAP: An Ecological Function Oriented Metagenomic Analysis Pipeline

Figure S 10 to **Figure S 14** for Chapter 5: A generalized Brody distribution based Random Matrix Theory approach for inferring microbial data association networks

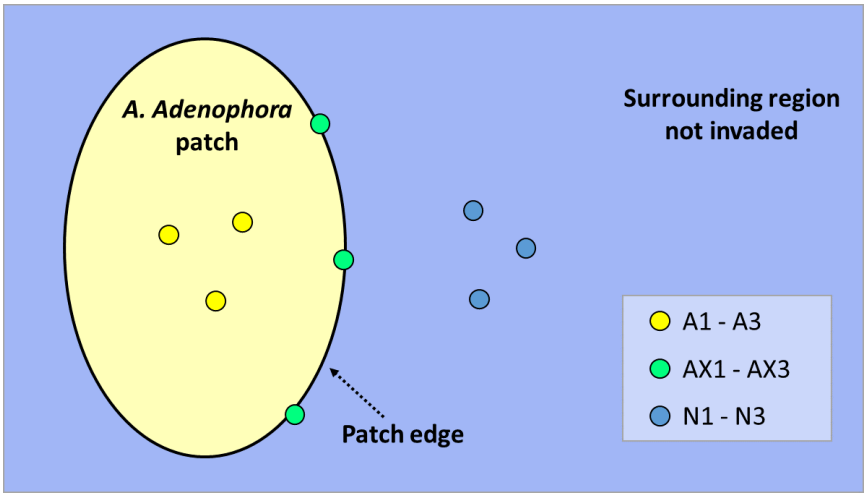


Figure S 1 Schematic diagram for the sampling site. Three samples (A1 -A3, yellow points) were collected from center of a patch (light yellow region) dominated by *A. adenophora*; three samples (AX1 -AX3, green points) were collected from a mix region (black border of the patch) around edge of the patch where *A. adenophora* and native plants co-existed; and three samples (N1 - N3, blue points) were collected from surrounding region of the patch where *A. adenophora* was absent. Each soil sample was collected from the remaining soil in the hole generated by removing randomly chosen plant and rhizosphere soil (30 cm radius around the plant). Samples were homogenized and sieved (2 mm) to remove stones, roots and soil animals.

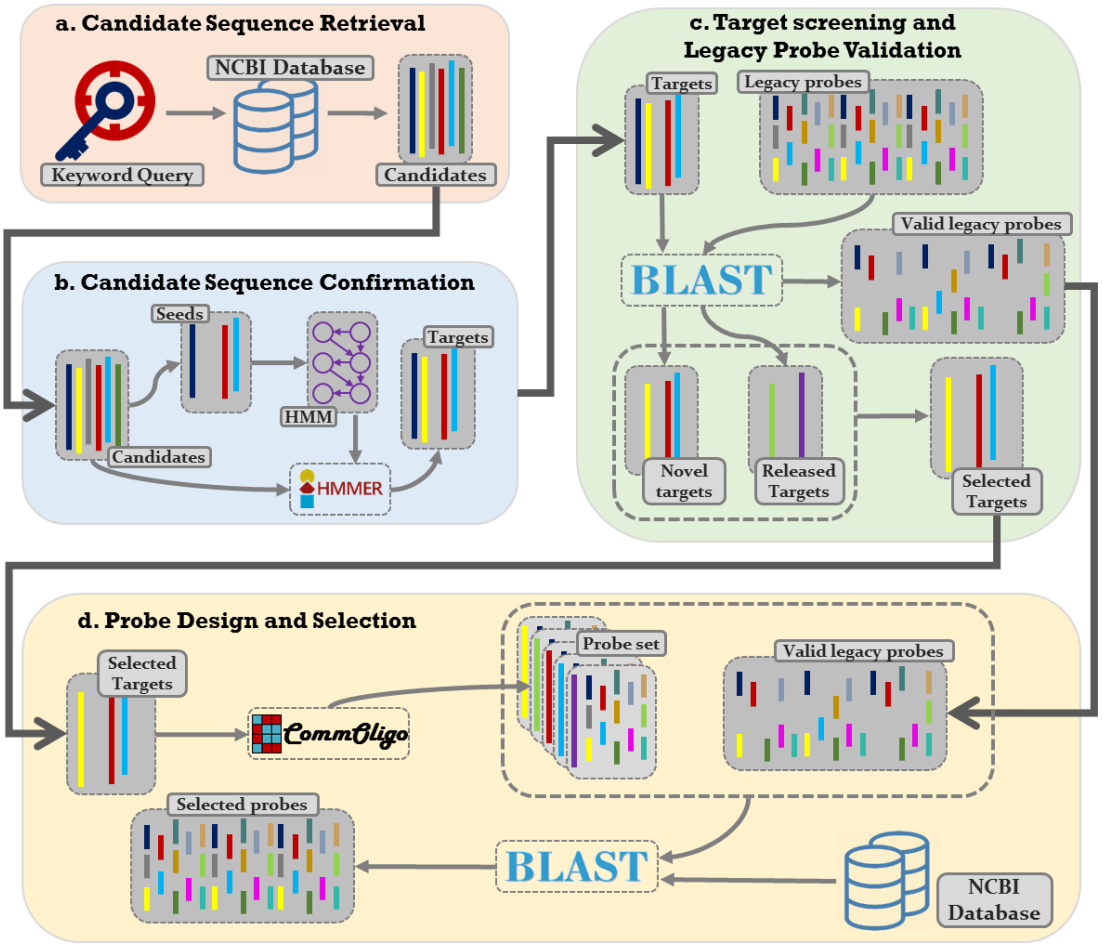


Figure S 2 The workflow of GeoChip 5.0 development started with the formulation and submission of keyword query for every functional gene family, then the retrieval of candidate sequences, which was followed candidate sequence confirmation, target screening and legacy probe validation, and probe design and selection.

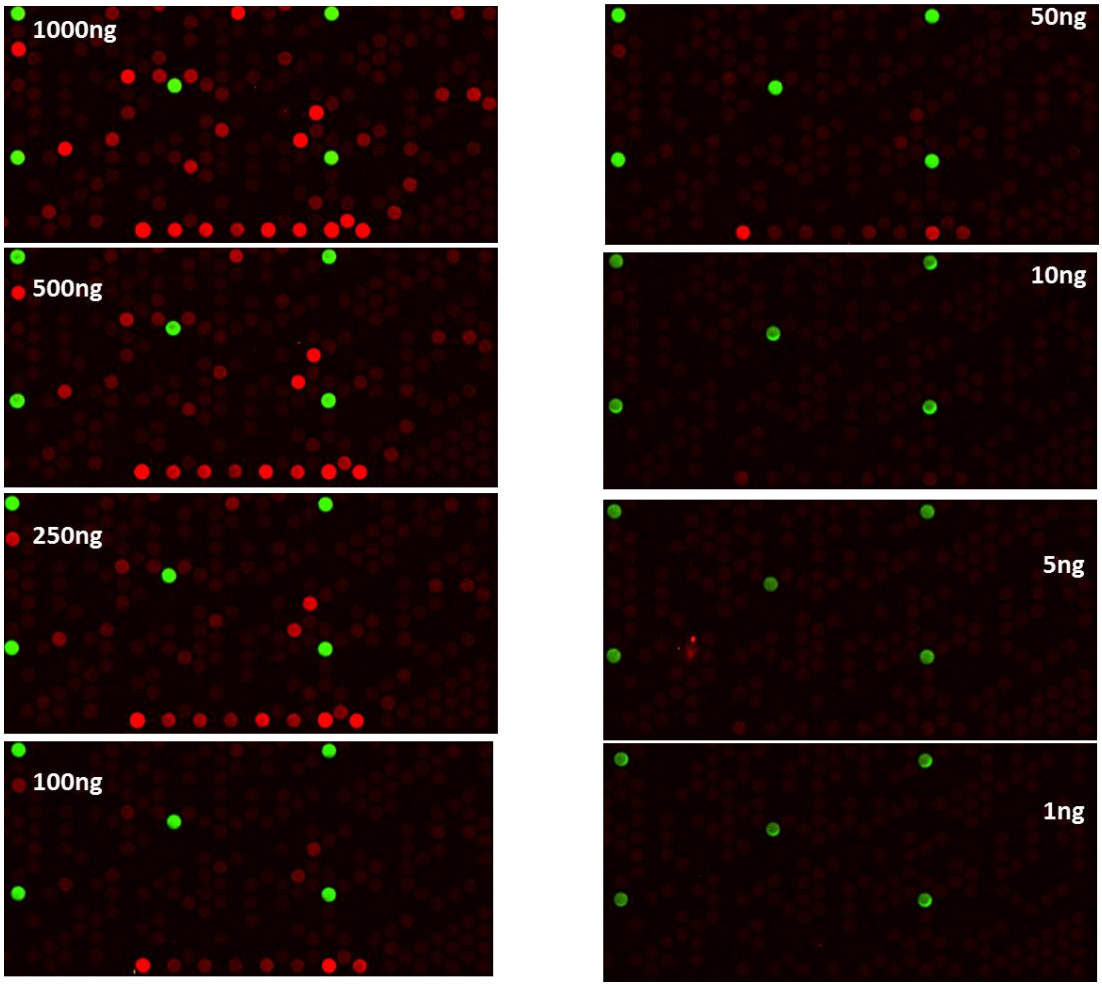


Figure S 3 Effects of DNA concentrations on hybridization. Different amounts of community DNAs from a grassland soil, ranging 1 ng to 1,000 ng was labeled with Cy5 (red spots) and hybridized to GeoChip 5.0S at 67 C, plus 10% formamide. Small amount of Cy3 labeled CORS was also added to the hybridization solution as control (green spots).

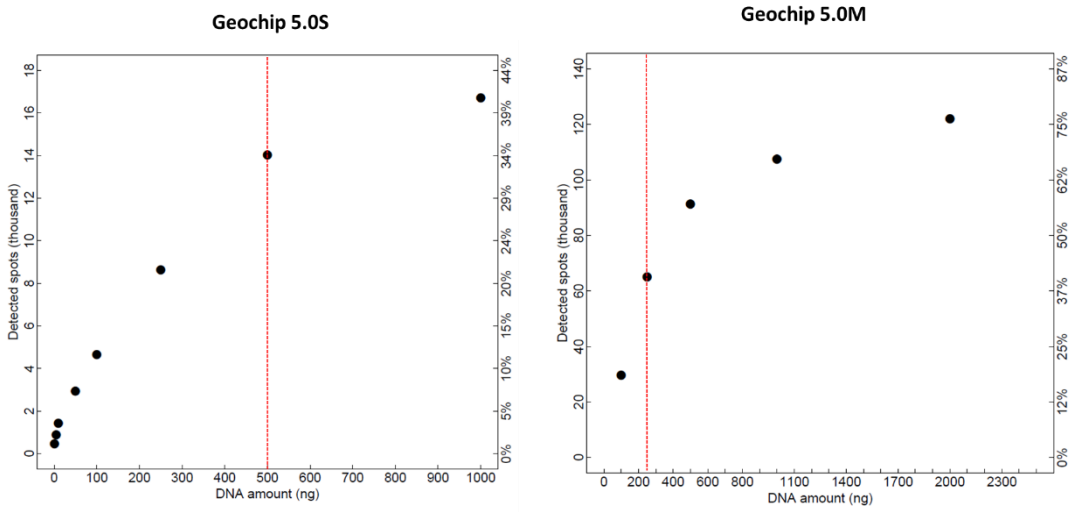


Figure S 4 Relationships between spots detected and the concentrations of community DNAs. (**a**) Hybridization of community DNAs from a grassland soils with GeoChip 5.0S (see the images in Fig S2). (**b**) Hybridization of community DNAs from a wastewater treatment plant with GeoChip 5.0M. Different amount of community DNAs were directly labeled with fluorescent dyes without any amplifications in triplicates and hybridizations were carried out at 67C plus 10% formamide for 24 hours. Red dashed lines show the DNA amounts for detecting more than 30% of spots in GeoChip 5.0S and 5.0M. Any spots with SNR > 2 were considered as positive spots.

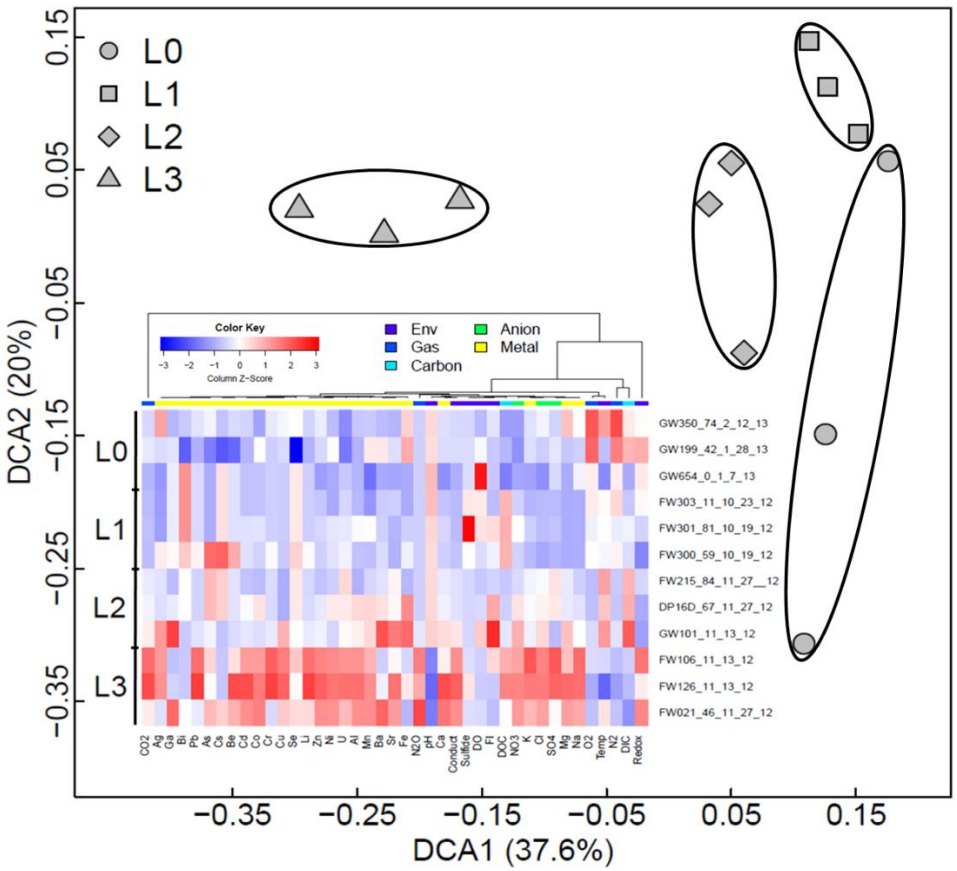


Figure S 5 Outer DCA showing on the difference of environmental factor values from 4 contamination levels (L0, L1, L2 and L3). Two axis (DCA1 and DCA2) were included the plot, which conveyed 37.6% and 20% explanatory power for the environmental factor difference. Inner heatmap showing the differences of 41 measured environmental factors across different samples from 4 contamination levels. The rows are the samples ordered in a top-down manner by the contamination levels from the lower to the higher. The columns are the measured environmental factors ordered as a form of dendrogram. The environmental factors are subjectively divided into five categories, including general environmental parameters (Env; details in **Table S 2**), gas TCD (Gas), organic matter (Carbon), anions (Anion) and metal ions (Metal), and marked with different color bars (purple, blue, cyan, green and yellow, respectively). Z-scores were calculated

based on original or log transformed values of each column, and rendered with a blue to red color gradient as Z-score value increased. The original values of conductivity, Cl, NO,3 SO4, Ag, Al, As, Ba, Be, Bi, Ca, Cd, Co, Cr, Cs, Cu, Fe, Ga, K, Li, Mg, Mn, Na, Ni, Pb, Se, Sr, U and Zn were log transformed due to the nature of the measurements.

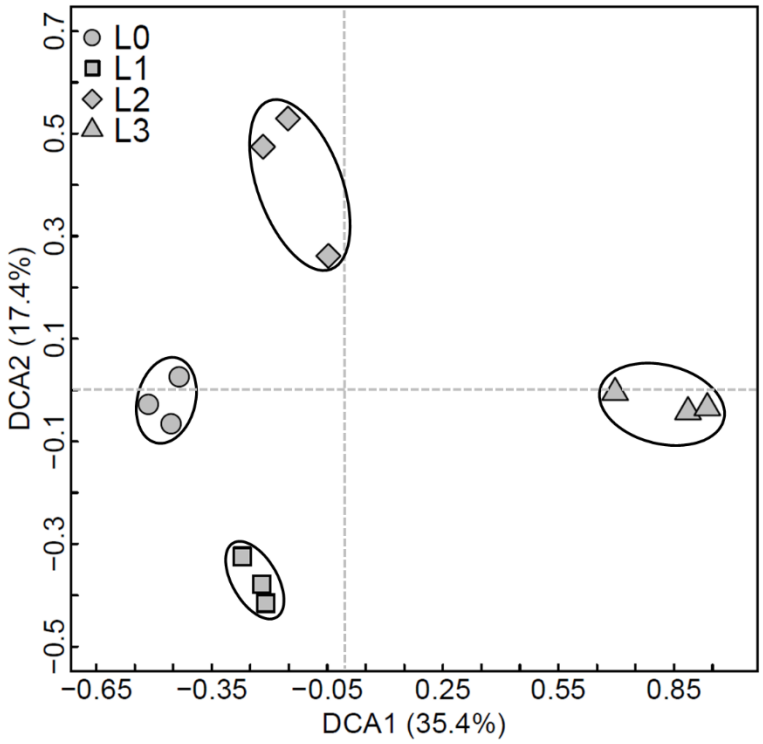


Figure S 6 DCA on the impacts of contaminants of different levels on the groundwater microbial community structure. Two axis (DCA1 and DCA2) were included the plot, which conveyed 35.4% and 17.4% explanatory power for the community structural difference.

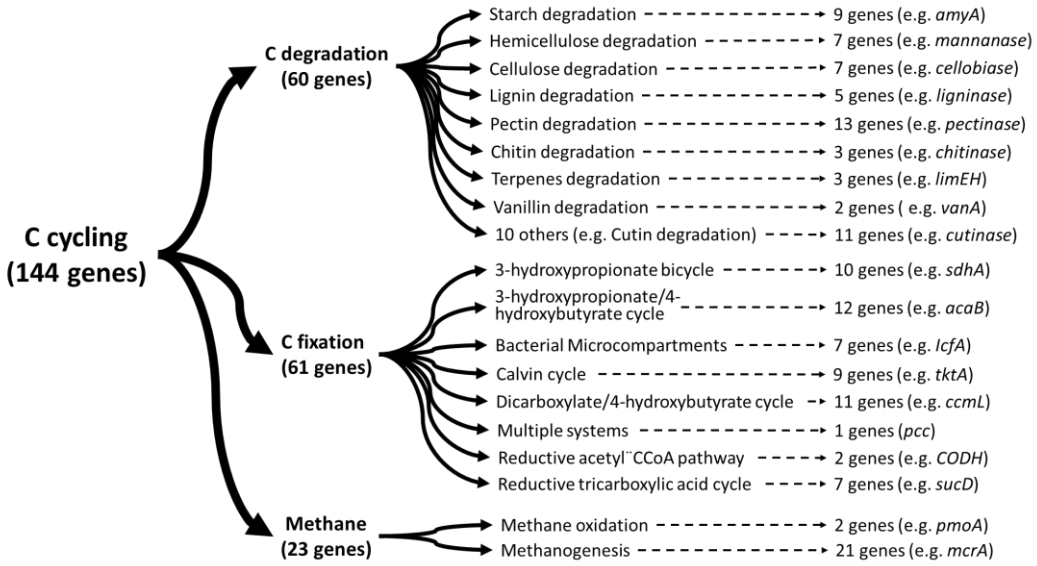


Figure S 7 An example of organization of functional genes in the EcoFun-MAP databases.

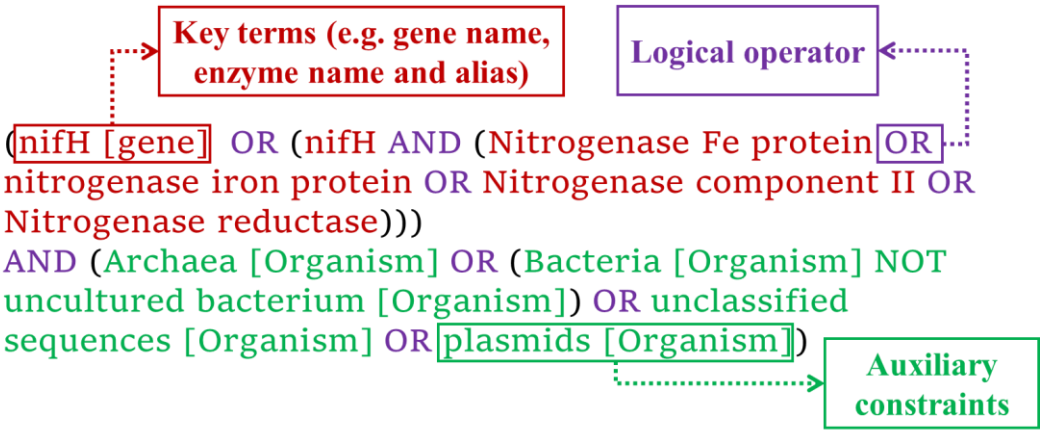


Figure S 8 An example showing components that constitute a typical keyword query.

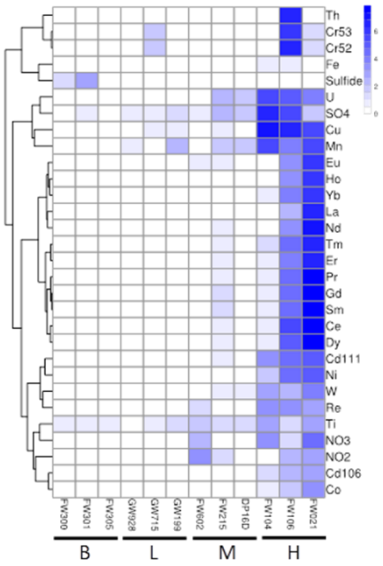


Figure S 9 Heatmap showing the levels of measurements of environmental factors among 12 underground water samples.

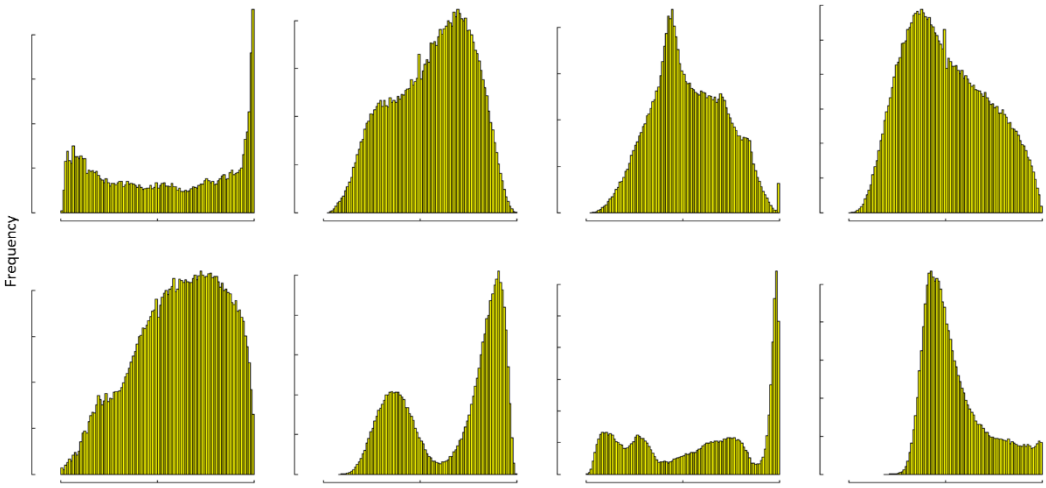


Figure S 10 Examples of irregular distributions of co-occurrence strengths from real datasets.

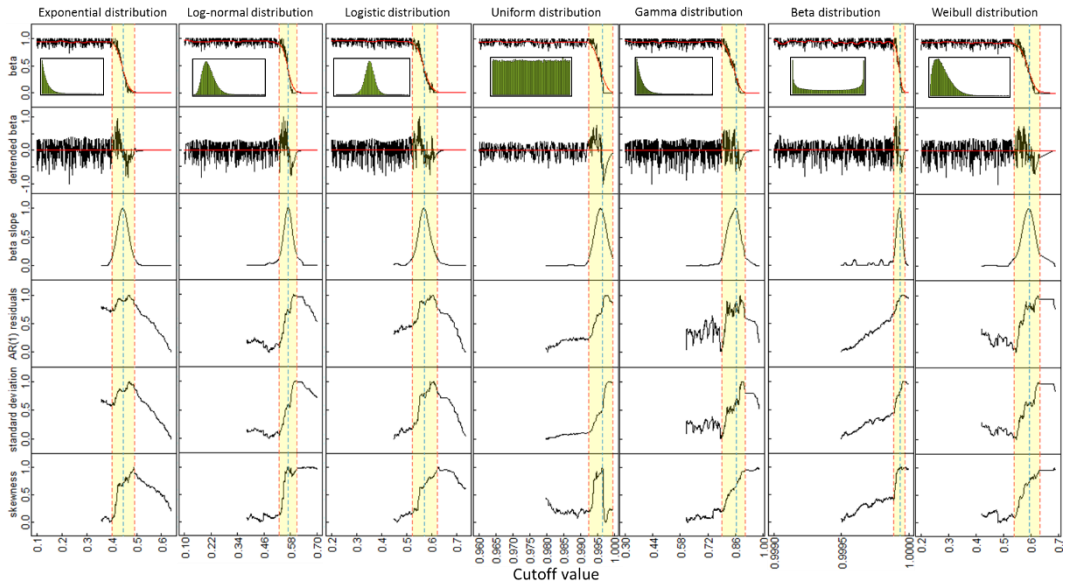


Figure S 11 Critical transitions in systems that are simulated based on different distributions**.**

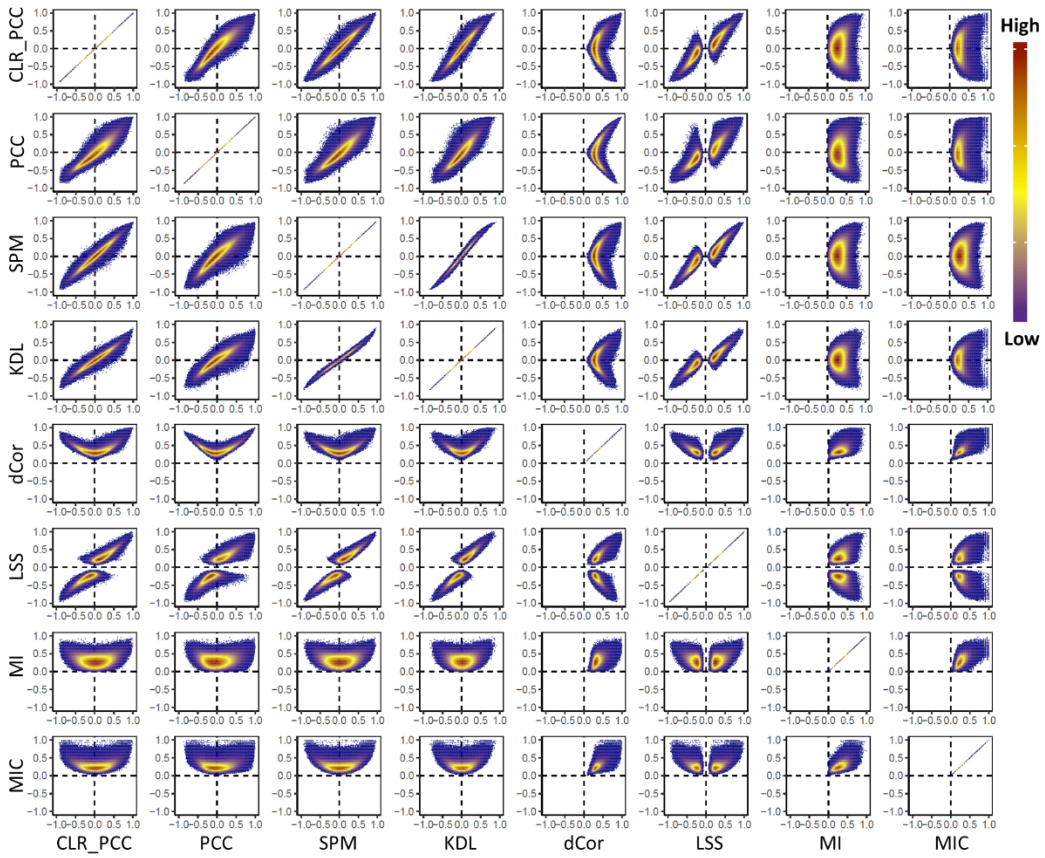


Figure S 12 Comparison of data association detection among different methods.

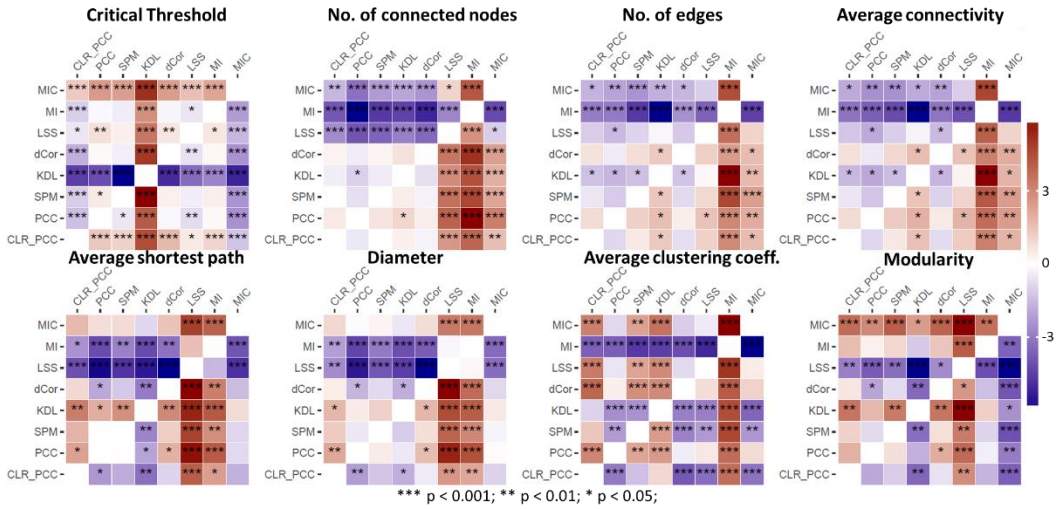


Figure S 13 Pairwise comparisons of detected thresholds and topological properties among the networks inferred using different data association detection methods for all datasets. The color and color depth of the heatmaps represent the signs and values of the Cohen's d. The levels of significance were determined by the paired t-test. The comparisons in the heatmaps were organized in a way that the methods of rows were compared to the methods of columns. For example, in the heatmap of critical threshold, the cell in MIC row and KDL column has deep red and three asteroids, it means critical threshold of network inferred using the MIC is significantly ($p < 0.001$) higher than the same using the KDL.

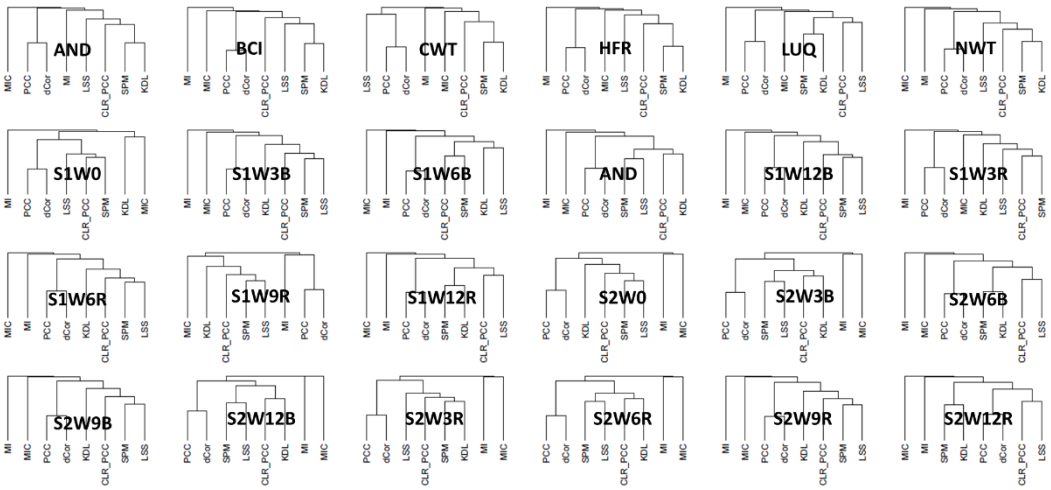


Figure S 14 Clustering different data association detection methods based on similarity of edges in the inferred networks for all datasets.

Appendix B: Supplementary Tables

Table S 1 and **Table S 2** for Chapter 3: Ultra-sensitive and -quantitative Detection of Microbial Populations in complex communities with New Functional Gene Arrays

Table S 3 and Table **Table S 4** for Chapter 4: The EcoFun-MAP: An Ecological Function Oriented Metagenomic Analysis Pipeline

Table S 5 and **Table S 6Figure S 14** for Chapter 5: A generalized Brody distribution based Random Matrix Theory approach for inferring microbial data association networks

Table S 1 Summary of probes in GeoChip 5.0M based on on phylogenetic distribution of the functional genes. B, bacteria; A, archaea; E, eukaryota; F, fungi; P, protist; V, virus; M, metazoan; U, unclassified. GS-probe, group-specific probe; SS-probe, sequence-specific group.

Domain-Kingdom-Phylum /order for virus	No. of										
	class	order	family	genus	species	strain	gene	covered CDS	probe	GS-probe	SS-probe
A-A-Crenarchaeota	2	6	9	22	43	57	142	13921	1312	1133	179
A-A-Euryarchaeota	10	12	23	69	133	187	229	33323	3992	3625	367
A-A-Korarchaeota	1	1	1	1	1	1	33	3718	35	33	2
A-A-Nanoarchaeota	1	1	1	1	1	1	6	12	6	6	0
A-A-Thaumarchaeota	1	4	5	7	9	14	46	5218	95	73	22
A-A-U	1	1	1	1	1	25	46	1885	288	115	173
B-B-Acidobacteria	5	4	4	8	13	18	155	20180	745	715	30
B-B-Actinobacteria	1	8	48	120	351	753	381	80060	15394	13618	1776
B-B-Aquificae	1	2	4	12	17	22	127	19379	533	498	35
B-B-Bacteroidetes	7	10	23	107	239	394	267	58970	9661	8790	871
B-B-Caldiserica	1	1	1	1	1	1	13	38	19	19	0
B-B-Candidatus poribacteria	1	1	1	1	1	1	2	5	4	1	3
B-B-Candidatus saccharibacteria	1	1	1	1	1	3	9	22	17	4	13
B-B-Chlamydiae	2	1	4	7	15	23	74	12652	241	230	11
B-B-Chlorobi	1	1	1	6	15	20	138	21932	995	944	51
B-B-Chloroflexi	8	11	12	14	19	28	209	32839	1353	1289	64
B-B-Chrysiogenetes	1	1	1	2	2	2	33	81	41	40	1
B-B-Cyanobacteria	2	8	12	57	97	203	237	46219	4727	4209	518
B-B-Deferribacteres	1	1	1	4	4	4	81	2583	165	154	11
B-B-Deinococcus-thermus	2	3	4	7	18	27	172	21352	1026	978	48
B-B-Dictyoglomi	1	1	1	1	2	2	51	9447	99	98	1
B-B-Elusimicrobia	2	2	2	2	2	2	34	5713	50	49	1
B-B-Fibrobacteres	2	1	1	1	1	2	50	6671	69	64	5
B-B-Firmicutes	6	10	42	176	593	1556	429	107993	25601	21921	3680
B-B-Fusobacteria	2	1	2	5	15	38	111	17980	541	493	48
B-B-Gemmatimonadetes	1	1	1	1	1	1	51	8772	71	69	2
B-B-Ignavibacteriae	1	1	2	2	2	2	27	85	43	42	1
B-B-Lentisphaerae	2	3	2	2	2	2	56	4823	93	91	2
B-B-Nitrospirae	2	2	2	4	5	10	114	8777	237	197	40
B-B-Planctomycetes	3	3	4	15	23	32	183	26590	994	765	229

B-B-Poribacteria	1	1	1	1	1	1	13	25	13	12	1
B-B-Proteobacteria	7	47	115	508	1132	2855	823	187705	68319	59749	8570
B-B-Spirochaetes	2	1	3	8	48	80	172	11471	1007	829	178
B-B-Synergistetes	2	2	2	11	13	14	109	11410	472	392	80
B-B-Tenericutes	1	3	4	7	39	82	56	11489	266	218	48
B-B-Thermodesulfobacteria	1	1	1	2	5	6	50	162	85	77	8
B-B-Thermotogae	1	1	2	8	19	26	93	16870	588	553	35
B-B-Verrucomicrobia	4	6	8	14	17	30	172	29149	947	824	123
B-B-U	1	2	2	7	8	239	233	33937	6737	2583	4154
E-F-Ascomycota	10	25	70	164	245	402	201	19120	6995	2471	4524
E-F-Basidiomycota	7	24	61	104	122	175	119	8732	1549	530	1019
E-F-Chytridiomycota	1	2	2	1	2	2	7	10	10	0	10
E-F-Glomeromycota	1	2	2	3	3	3	9	71	12	7	5
E-F-Microsporidia	1	1	7	9	14	16	25	1900	82	47	35
E-F-Neocallimastigomycota	1	1	1	3	3	8	7	136	28	4	24
E-F-U	1	4	10	13	15	20	19	725	180	90	90
E-P-Apicomplexa	2	3	6	7	19	47	44	1875	374	123	251
E-P-Bacillariophyta	5	10	14	16	19	31	37	1269	218	96	122
E-P-Chromerida	1	1	1	2	2	2	3	6	5	1	4
E-P-Euglenida	1	4	6	6	6	7	3	17	12	4	8
E-P-Eustigmatophyceae	1	1	1	1	1	1	1	1	1	0	1
E-P-Haplosporidia	1	1	2	4	4	5	1	12	6	3	3
E-P-Phaeophyceae	1	2	2	2	2	2	19	1129	28	6	22
E-P-Pinguiophyceae	1	1	1	1	1	1	1	1	1	0	1
E-P-Xanthophyceae	1	2	2	2	2	2	4	7	5	2	3
E-M-Arthropoda	2	5	5	7	7	7	4	19	13	5	8
E-M-Chordata	2	3	3	3	3	3	3	9	7	2	5
E-M-Echinodermata	1	1	1	1	1	1	1	2	3	1	2
E-M-Nematoda	1	2	3	3	3	4	6	9	8	2	6
E-M-Platyhelminthes	1	1	1	1	1	1	1	5	3	2	1
E-Viridiplantae-Chlorophyta	7	15	23	37	39	57	39	694	444	221	223
E-Viridiplantae-Streptophyta	7	11	12	12	12	13	15	432	31	12	19
E-U-U	24	74	116	178	195	264	165	3148	1401	569	832
V-V-Nidovirales	-	-	1	7	18	75	2	259	143	60	83
V-V-Picornavirales	-	-	6	24	47	275	7	707	518	144	374
V-V-Tymovirales	-	-	1	6	8	46	2	236	111	79	32
V-V-Caudovirales	-	-	4	26	30	156	40	1506	347	279	68
V-V-U	-	-	28	104	208	814	78	3320	1729	768	961
U--	-	-	-	-	-	33	125	2561	816	293	523

Table S 2 Measurements of environmental variables from the underground water samples used in the application study. A total of 12 samples were collected from different wells that were contaminated at 4 different levels (L0, L1, L2 and L3). Measurements included 5 major categories: general environmental parameters (Env. Parameters), gas TCD, dissolved Carbon (C), anion and metal ion.

		L0			L1			L2			L3		
		GW-654	GW-199	GW-350	FW-300	FW-301	FW-303	FW-215	DP16D	GW-101	FW-106	FW-126	FW-021
General Env. Parameters	Temp. (°C)	13.0	14.3	18.0	15.5	15.8	15.8	17.9	17.1	18.1	14.8	12.2	16.4
	D.O. (mg/L)	2.4	0.5	0.02	0.3	0.8	0.7	0.2	0.3	1.0	0.2	0.2	0.3
	Cdt. (μS/cm)	269	582	545	379	334	316	637	661	1721	7864	18770	7967
	Redox (mV)	175	305	137	-129	39	147	43	-50	-72	426	168	387
	pH	7.19	6.53	6.67	6.59	6.68	7.16	6.6	6.67	6.81	3.55	3.04	3.43
	S^{2-} (mg/L)	0.003	0.003	0	0.026	0.188	0.041	0	0	0	0.004	0.044	0.004
	F.I. (mg/L)	0.1	0.13	0.38	0.5	0.98	0	0.21	1.46	2.7	1.02	0.13	0.03
Gas TCD (μmol/mL)	N_2	59	104	126	57*	46	69	32	31	43	28	14	30
	O_2	12	23	25	11*	10	13	6	6	8	8	4	6
	CO_2	37	172	128	54*	60	48	173	202	283	631	718	289
	N_2O	0	0	0	0	0	0	0	0	2.7	22.6	17.0	30.1
C (mg/L)	D.I.C	30.98	85.05	67.46	48.13	55.44	40.82	85.25	87.91	115.9	43.18	36.65	22.27
	D.O.C.	0.345	1.335	0.717	44.54	48.65	39.59	1.928	2.326	4.065	47.87	128.2	7.298
Anion (mg/L)	Cl	1.3	5.5	13.5	2.2	3.5	3.3	15.6	23.7	42.2	318.2	373.7	152.3
	NO_3^+	0.5	0.2	0.3*	3.7	36.4	4.0	5.5	141.0	1470.9	2692	11648	4507
	SO_4^+	16	21	13	6	9	7	76	65	8	2063	1460	42
Metal (mg/L)	Ag	0.007	0.010	0.021	0.014	0.008	0.008	0.011	0.011	0.022	0.022	0.023	0.011
	Al	0.015	0.013	0.038	0.03	0.418	0.017	0.013	3.444	1.129	108	559	115
	As	0.003	0.002	0.005	0.021	0.003	0.003	0.011	0.011	0.005	0.008	0.006	0.011
	Ba	0.04	0.242	0.055	0.093	0.077	0.072	0.104	0.269	2.820	0.077	0.131	2.073
	Be	0.04	0.019	0.038	0.082	0.04	0.04	0.041	0.041	0.039	0.059	0.149	0.079
	Bi	0.038	0.001	0.005	0.018	0.038	0.038	0.009	0.009	0.005	0.005	0.005	0.009
	Ca	17	89	92	67	79	53	120	140	420	273	9838	3970
	Cd	0.003	0.002	0.004	0.006	0.006	0.003	0.003	0.01	0.007	0.132	0.866	0.173
	Co	0.004	0.000	0.002	0.006	0.011	0.004	0.003	0.056	0.009	0.509	1.364	1.225
	Cr	0.004	0.004	0.007	0.010	0.004	0.004	0.005	0.005	0.007	0.384	0.798	0.005
	Cs	0.029	0.008	0.016	0.064	0.029	0.029	0.032	0.032	0.016	0.017	0.016	0.032
	Cu	0.009	0.013	0.025	0.008	0.009	0.009	0.004	0.004	0.222	0.810	1.587	0.118
	Fe	0.011	0.851	0.338	0.067	0.030	0.011	0.016	1.933	4.585	0.038	0.167	0.016
	Ga	0.011	0.010	0.007	0.014	0.011	0.011	0.007	0.011	0.089	0.007	0.009	0.06
	K	1.0	3.3	3.4	1.5	2.4	1.5	5.2	6.0	4.9	216.4	102.5	28.8
	Li	0.038	0.022	0.042	0.096	0.038	0.038	0.048	0.048	0.069	1.946	5.190	0.227
	Mg	16.2	22.3	80.9	16.2	16.2	16.2	32.3	32.3	80.9	80.9	216.2	117.9
	Mn	0.0	4.3	0.3	0.2	2.2	0.0	0.5	9.4	8.3	32.5	134.1	128.5
	Na	39.9	20.8	52.0	10.4	10.4	10.4	20.8	29.3	52.0	865.3	826.3	269.2
	Ni	0.006	0	0.011	0.043	0.048	0.006	0.021	0.242	0.026	7.184	15.352	5.339
	Pb	0.003	0.002	0.003	0.006	0.003	0.003	0.003	0.003	0.003	0.032	0.060	0.004
	Se	0.009	0	0.004	0.018	0.009	0.009	0.009	0.009	0.004	0.024	0.005	0.013
	Sr	0.096	0.212	0.119	0.115	0.12	0.168	0.399	0.372	2.219	0.369	2.43	1.373
	U	0.051	0.003	0.006	0.216	0.16	0.081	1.452	0.744	0.417	16.625	55.286	3.751
	Zn	0.02	0.059	0.084	0.051	0.058	0.041	0.040	0.078	0.093	1.099	2.189	0.897

* missing values were imputed using the mean of values from other two replicates in the same group.

Table S 3 General information about 12 underground water samples, and results of HiSeq shotgun metagenomics sequencing and the EcoFun-MAP analyses

Sample ID	Group label	Contamination level	No. of HiSeq reads (M)/ data amount (Gbp)	No. of hits (M)/percentage (%)				
				Ultra-fast	Fast	Moderate	Sensitive	Ultra-sensitive
FW300	L0	Background	~266/39.9	~2.6/0.99	~8.9/3.35	~0.2/0.06	~0.2/0.07	~0.2/0.08
FW301	L0	Background	~164.5/25.7	~1.8/1.08	~5.9/3.62	~0.1/0.08	~0.2/0.09	~0.2/0.09
FW305	L0	Background	~97.2/14.6	~1.3/1.30	~4.5/4.63	~0.1/0.14	~0.2/0.18	~0.2/0.18
GW199	L1	Low	~79.6/11.9	~1.1/1.35	~3.5/4.41	~0.1/0.16	~0.2/0.19	~0.1/0.18
GW928	L1	Low	~92.5/13.9	~1.3/1.36	~4.3/4.61	~0.1/0.13	~0.1/0.16	~0.1/0.16
GW715	L1	Low	~202.3/30.3	~2.7/1.34	~9.6/4.76	~0.2/0.12	~0.3/0.16	~0.3/0.16
DP16D	L2	Medium	~195.9/29.4	~2.4/1.23	~7.7/3.92	~0.1/0.07	~0.2/0.08	~0.2/0.08
FW215	L2	Medium	~171.3/25.7	~1.9/1.09	~6.0/3.52	~0.1/0.06	~0.1/0.07	~0.1/0.07
FW602	L2	Medium	~154.1/23.1	~2.0/1.29	~6.9/4.45	~0.2/0.12	~0.2/0.15	~0.2/0.15
FW104	L3	High	~94.5/14.2	~1.3/1.41	~4.8/5.05	~0.2/0.18	~0.2/0.21	~0.2/0.22
FW106	L3	High	~119.6/17.9	~2.1/1.74	~7.9/6.58	~0.3/0.27	~0.4/0.34	~0.4/0.35
FW021	L3	High	~178.9/26.8	~3.0/1.69	~11.1/6.20	~0.3/0.18	~0.4/0.22	~0.4/0.23
Total	-	-	~1816.7/272.5	~23.4/1.29	~81.1/4.46	~2.1/0.12	~2.7/0.15	~2.7/0.15

Table S 4 Summary of results for evaluating accuracy and precision of five workflows in the EcoFun-MAP. The results here are based on counts of hits from the running of five workflows on each sample.

Mode	Sample ID	Accuracy rate (%)				Precision rate (%)			
		Level 1	Level 2	Level 3	Level 4	Level 1	Level 2	Level 3	Level 4
ultra-fast	FW300	71.5	74.0	74.9	78.2	5.5	5.7	5.8	6.0
	FW301	72.5	74.7	75.6	78.8	6.2	6.4	6.4	6.7
	FW305	68.8	70.8	71.7	74.7	9.5	9.8	9.9	10.4
	GW199	72.9	74.7	75.2	77.8	9.6	9.8	9.9	10.2
	GW928	70.1	72.1	72.9	76.0	8.2	8.5	8.6	8.9
	GW715	67.1	69.3	70.2	73.9	8.2	8.4	8.5	9.0
	DP16D	70.9	73.3	73.8	76.3	4.8	5.0	5.0	5.2
	FW215	73.5	76.1	76.7	79.7	4.9	5.0	5.1	5.3
	FW602	70.9	73.5	74.3	77.8	8.4	8.7	8.8	9.2
	FW104	73.4	77.7	78.0	80.6	11.3	12.0	12.0	12.4
	FW106	68.3	74.4	74.7	77.6	13.8	15.0	15.0	15.6
	FW021	69.6	75.5	76.3	79.9	9.4	10.1	10.3	10.7
fast	FW300	85.1	87.4	87.8	91.2	1.9	2.0	2.0	2.1
	FW301	86.4	88.4	88.9	92.0	2.2	2.2	2.3	2.3
	FW305	86.8	88.8	89.3	92.4	3.4	3.5	3.5	3.6
	GW199	87.3	88.6	88.9	91.1	3.5	3.6	3.6	3.7
	GW928	86.6	88.3	88.7	91.8	3.0	3.1	3.1	3.2
	GW715	85.8	87.9	88.4	92.1	2.9	3.0	3.0	3.2
	DP16D	85.0	86.7	86.8	89.4	1.8	1.8	1.8	1.9
	FW215	86.0	88.1	88.4	91.6	1.8	1.8	1.8	1.9
	FW602	84.9	86.9	87.4	91.0	2.9	3.0	3.0	3.1
	FW104	86.7	88.4	88.8	91.6	3.7	3.8	3.8	3.9
	FW106	83.9	88.9	89.6	92.8	4.5	4.7	4.8	4.9
	FW021	84.3	89.4	90.2	93.2	3.1	3.3	3.3	3.4
moderate	FW300	69.0	69.2	69.2	69.5	86.1	86.3	86.3	86.7
	FW301	70.1	70.2	70.2	70.5	84.1	84.3	84.3	84.6
	FW305	66.5	66.7	66.7	67.0	86.2	86.5	86.5	86.9
	GW199	71.1	71.2	71.2	71.5	78.3	78.5	78.5	78.8
	GW928	68.0	68.1	68.1	68.5	84.7	84.8	84.8	85.3
	GW715	65.3	65.4	65.5	65.8	88.0	88.2	88.2	88.7
	DP16D	69.0	69.1	69.1	69.3	82.4	82.6	82.6	82.8
	FW215	71.5	71.7	71.7	72.0	82.9	83.1	83.1	83.4
	FW602	69.6	69.7	69.8	70.1	88.7	88.9	88.9	89.3
	FW104	73.2	73.3	73.3	73.5	87.1	87.2	87.2	87.4
	FW106	69.4	69.6	69.6	69.8	91.1	91.4	91.4	91.7
	FW021	70.8	71.0	71.0	71.3	90.5	90.7	90.8	91.1
sensitive	FW300	82.9	83.1	83.1	83.4	85.2	85.3	85.3	85.7
	FW301	84.4	84.5	84.5	84.8	84.1	84.2	84.2	84.5
	FW305	84.2	84.4	84.4	84.7	84.7	84.9	85.0	85.3
	GW199	84.2	84.3	84.3	84.5	79.0	79.1	79.1	79.4
	GW928	83.8	83.9	83.9	84.3	84.0	84.1	84.1	84.4
	GW715	83.6	83.8	83.8	84.2	85.6	85.8	85.8	86.2
	DP16D	82.5	82.6	82.6	82.8	83.1	83.2	83.2	83.5
	FW215	84.0	84.2	84.2	84.5	83.3	83.5	83.5	83.7
	FW602	84.4	84.5	84.5	84.9	87.2	87.4	87.4	87.7
	FW104	85.6	85.7	85.7	85.9	86.7	86.8	86.8	87.0
	FW106	86.8	86.9	87.0	87.2	88.7	88.9	88.9	89.1
	FW021	86.3	86.4	86.4	86.7	88.7	88.9	88.9	89.2

Table S 5 Basic information about 16S profiling datasets from two real projects.

Sample ID	Project	Data type	No. of all replicates	No. of min. presences in replicates (> 60%)	No. of OTUs
AND	Biogeographic survey	16S	21	13	1,880
BCI	Biogeographic survey	16S	21	13	2,077
CWT	Biogeographic survey	16S	21	13	1,804
HFR	Biogeographic survey	16S	21	13	1,479
LUQ	Biogeographic survey	16S	21	13	1,713
NWT	Biogeographic survey	16S	21	13	1,542
S1W0	Plant succession	16S	16	10	1,870
S1W3B	Plant succession	16S	16	10	1,925
S1W6B	Plant succession	16S	16	10	1,952
S1W9B	Plant succession	16S	16	10	1,939
S1W12B	Plant succession	16S	16	10	1,906
S1W3R	Plant succession	16S	16	10	1,773
S1W6R	Plant succession	16S	16	10	1,676
S1W9R	Plant succession	16S	16	10	1,655
S1W12R	Plant succession	16S	16	10	1,466
S2W0	Plant succession	16S	16	10	2,039
S2W3B	Plant succession	16S	16	10	2,074
S2W6B	Plant succession	16S	16	10	2,108
S2W9B	Plant succession	16S	16	10	2,113
S2W12B	Plant succession	16S	16	10	2,132
S2W3R	Plant succession	16S	16	10	1,933
S2W6R	Plant succession	16S	16	10	1,805
S2W9R	Plant succession	16S	16	10	1,746
S2W12R	Plant succession	16S	16	10	1,634

Table S 6 Detailed information about the detected critical transitions and thresholds in all datasets based on different data association methods.

Sample ID	Key thresholds (Beginning/ending of transition/final threshold)																							
	CLR_PCC			PCC			SPM			KDL			dCor			LSS			MI			MIC		
AND	0.746	0.932	0.903	0.744	0.889	0.844	0.699	0.883	0.855	0.568	0.756	0.735	0.704	0.893	0.867	0.737	0.904	0.889	0.673	0.836	0.808	0.758	0.996	0.962
BCI	0.671	0.917	0.901	0.644	0.929	0.876	0.654	0.849	0.84	0.523	0.727	0.705	0.681	0.871	0.842	0.646	0.871	0.858	0.717	0.84	0.815	0.611	0.997	0.872
CWT	0.702	0.937	0.924	0.748	0.911	0.826	0.702	0.893	0.862	0.566	0.774	0.737	0.732	0.874	0.838	0.705	0.877	0.82	0.69	0.824	0.808	0.735	0.997	0.923
HFR	0.672	0.866	0.796	0.676	0.882	0.788	0.682	0.847	0.813	0.569	0.714	0.68	0.68	0.855	0.761	0.662	0.876	0.86	0.732	0.816	0.801	0.682	0.917	0.868
LUQ	0.671	0.867	0.855	0.657	0.852	0.787	0.634	0.827	0.768	0.517	0.682	0.67	0.637	0.839	0.815	0.671	0.867	0.856	0.729	0.828	0.799	0.645	0.918	0.828
NWT	0.709	0.883	0.861	0.717	0.867	0.817	0.679	0.858	0.831	0.513	0.724	0.685	0.712	0.857	0.818	0.691	0.862	0.847	0.666	0.768	0.732	0.739	0.975	0.922
S1W0	0.745	0.833	0.784	0.707	0.789	0.76	0.719	0.797	0.765	0.627	0.706	0.683	0.714	0.793	0.764	0.698	0.8	0.745	0.63	0.752	0.749	0.756	0.885	0.822
S1W3B	0.739	0.825	0.803	0.707	0.798	0.754	0.694	0.822	0.775	0.62	0.708	0.676	0.715	0.796	0.757	0.708	0.818	0.772	0.661	0.814	0.78	0.773	0.917	0.83
S1W6B	0.711	0.871	0.773	0.696	0.859	0.799	0.702	0.838	0.765	0.605	0.737	0.677	0.702	0.848	0.775	0.654	0.849	0.818	0.568	0.793	0.74	0.648	0.985	0.9
S1W9B	0.758	0.836	0.814	0.709	0.779	0.754	0.72	0.798	0.772	0.627	0.723	0.669	0.715	0.792	0.763	0.734	0.84	0.793	0.645	0.792	0.78	0.781	0.889	0.83
S1W12B	0.74	0.833	0.787	0.704	0.791	0.752	0.715	0.802	0.769	0.621	0.71	0.673	0.71	0.789	0.766	0.688	0.788	0.755	0.647	0.792	0.767	0.745	0.888	0.819
S1W3R	0.735	0.812	0.791	0.71	0.795	0.758	0.721	0.796	0.776	0.629	0.708	0.685	0.713	0.79	0.769	0.745	0.842	0.811	0.651	0.773	0.773	0.749	0.865	0.83
S1W6R	0.722	0.854	0.789	0.696	0.82	0.753	0.715	0.854	0.77	0.621	0.718	0.668	0.713	0.812	0.764	0.658	0.792	0.773	0.622	0.747	0.728	0.705	0.927	0.897
S1W9R	0.74	0.825	0.796	0.709	0.798	0.759	0.723	0.805	0.78	0.604	0.747	0.686	0.717	0.793	0.772	0.727	0.835	0.797	0.589	0.703	0.689	0.789	0.915	0.832
S1W12R	0.737	0.858	0.785	0.698	0.835	0.768	0.659	0.872	0.788	0.624	0.731	0.676	0.705	0.814	0.769	0.662	0.793	0.79	0.655	0.745	0.737	0.751	0.928	0.897
S2W0	0.755	0.854	0.811	0.711	0.794	0.76	0.721	0.805	0.77	0.626	0.711	0.681	0.718	0.802	0.776	0.701	0.805	0.759	0.701	0.847	0.803	0.779	0.927	0.898
S2W3B	0.767	0.847	0.823	0.712	0.796	0.766	0.725	0.802	0.772	0.628	0.712	0.679	0.72	0.796	0.769	0.746	0.849	0.798	0.677	0.825	0.802	0.79	0.927	0.899
S2W6B	0.717	0.849	0.769	0.695	0.841	0.78	0.711	0.826	0.796	0.611	0.721	0.669	0.711	0.818	0.762	0.69	0.845	0.763	0.619	0.848	0.812	0.795	0.982	0.901
S2W9B	0.766	0.877	0.819	0.713	0.809	0.767	0.723	0.819	0.782	0.624	0.725	0.691	0.718	0.817	0.766	0.719	0.845	0.78	0.728	0.83	0.81	0.798	0.928	0.9
S2W12B	0.773	0.861	0.823	0.708	0.806	0.759	0.722	0.809	0.768	0.618	0.708	0.679	0.718	0.8	0.783	0.696	0.819	0.753	0.712	0.787	0.753	0.782	0.928	0.904
S2W3R	0.745	0.847	0.821	0.71	0.799	0.763	0.726	0.808	0.78	0.624	0.717	0.681	0.719	0.806	0.764	0.744	0.863	0.812	0.653	0.771	0.747	0.797	0.927	0.898
S2W6R	0.769	0.855	0.82	0.707	0.801	0.762	0.716	0.815	0.763	0.619	0.701	0.677	0.715	0.801	0.759	0.727	0.868	0.803	0.675	0.814	0.771	0.785	0.927	0.899
S2W9R	0.737	0.826	0.788	0.709	0.805	0.78	0.721	0.811	0.775	0.626	0.715	0.68	0.718	0.808	0.761	0.737	0.854	0.793	0.715	0.838	0.803	0.791	0.928	0.901
S2W12R	0.73	0.864	0.78	0.703	0.823	0.765	0.69	0.851	0.776	0.617	0.714	0.672	0.714	0.821	0.772	0.669	0.8	0.746	0.708	0.803	0.76	0.79	0.988	0.957

Reference

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

Aitchison, J. (1986). "The statistical analysis of compositional data."

Albert, R. (2005). "Scale-free networks in cell biology." Journal of cell science **118**(21): 4947-4957.

Albert, R., H. Jeong and A.-L. Barabási (2000). "Error and attack tolerance of complex networks." Nature **406**(6794): 378-382.

Algora, C., S. Vasileiadis, K. Wasmund, M. Trevisan, M. Krüger, E. Puglisi and L. Adrian (2015). "Manganese and iron as structuring parameters of microbial communities in Arctic marine sediments from the Baffin Bay." FEMS Microbiology Ecology **91**(6).

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.

Altschul, S. F., T. L. Madden, A. A. Schäffer, J. Zhang, Z. Zhang, W. Miller and D. J. Lipman (1997). "Gapped BLAST and PSI-BLAST: a new generation of protein database search programs." Nucleic Acids Research **25**(17): 3389-3402.

Anderson, M. J. (2001). "A new method for non- parametric multivariate analysis of variance." Austral Ecology **26**(1): 32-46.

Anderson, M. J. (2006). "Distance-based tests for homogeneity of multivariate dispersions." Biometrics **62**(1): 245-253.

Andrews, S. (2010). "FastQC: A quality control tool for high throughput sequence data." Reference Source.

Atlas, R. M. and R. Bartha (1986). "Microbial ecology: fundamentals and applications."

Bandyopadhyay, J. N. and S. Jalan (2007). "Universality in complex networks: Random matrix analysis." Physical Review E **76**(2): 026109.

Barabási, A.-L. and R. Albert (1999). "Emergence of scaling in random networks." Science **286**(5439): 509-512.

Barabási, A.-L. and Z. N. Oltvai (2004). "Network biology: understanding the cell's functional organization." Nature Reviews Genetics **5**(2): 101-113.

Barberán, A., S. T. Bates, E. O. Casamayor and N. Fierer (2012). "Using network analysis to explore co-occurrence patterns in soil microbial communities." The ISME journal **6**(2): 343-351.

Barka, E. A. and C. Clément (2008). Plant-microbe interactions, Research Signpost.

Barrett, M. T., A. Scheffer, A. Ben-Dor, N. Sampas, D. Lipson, R. Kincaid, P. Tsang, B. Curry, K. Baird and P. S. Meltzer (2004). "Comparative genomic hybridization using oligonucleotide microarrays and total genomic DNA." Proceedings of the National Academy of Sciences **101**(51): 17765-17770.

Batten, K. M., K. M. Scow, K. F. Davies and S. P. Harrison (2006). "Two invasive plants alter soil microbial community composition in serpentine grasslands." Biological Invasions **8**(2): 217-230.

Berendsen, R. L., C. M. J. Pieterse and P. A. H. M. Bakker (2012). "The rhizosphere microbiome and plant health." Trends in plant science **17**(8): 478-486.

Bossier, P., M. Hofte and W. Verstraete (1988). Ecological significance of siderophores in soil. Advances in microbial ecology, Springer**:** 385-414.

Brodie, E. L., T. Z. DeSantis, J. P. M. Parker, I. X. Zubietta, Y. M. Piceno and G. L. Andersen (2007). "Urban aerosols harbor diverse and dynamic bacterial populations." Proceedings of the National Academy of Sciences **104**(1): 299-304.

Broz, A. K., D. K. Manter and J. M. Vivanco (2007). "Soil fungal abundance and diversity: another victim of the invasive plant Centaurea maculosa." The ISME journal **1**(8): 763-765.

Buchfink, B., C. Xie and D. H. Huson (2015). "Fast and sensitive protein alignment using DIAMOND." Nature methods **12**(1): 59-60.

Callaway, R. M. and E. T. Aschehoug (2000). "Invasive plants versus their new and old neighbors: a mechanism for exotic invasion." Science **290**(5491): 521-523.

Callaway, R. M., D. Cipollini, K. Barto, G. C. Thelen, S. G. Hallett, D. Prati, K. Stinson and J. Klironomos (2008). "Novel weapons: invasive plant suppresses fungal mutualists in America but not in its native Europe." Ecology **89**(4): 1043-1055.

Callaway, R. M., G. C. Thelen, A. Rodriguez and W. E. Holben (2004). "Soil biota and exotic plant invasion." Nature **427**(6976): 731-733.

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

Carey, C. J., J. M. Beman, V. T. Eviner, C. M. Malmstrom and S. C. Hart (2015). "Soil microbial community structure is unaltered by plant invasion, vegetation clipping, and

nitrogen fertilization in experimental semi-arid grasslands." Frontiers in Microbiology **6**.

Chaffron, S., H. Rehrauer, J. Pernthaler and C. von Mering (2010). "A global network of coexisting microbes from environmental and whole-genome sequence data." Genome research **20**(7): 947-959.

Chaparro, J. M., D. V. Badri and J. M. Vivanco (2014). "Rhizosphere microbiome assemblage is affected by plant development." The ISME journal **8**(4): 790-803.

Cipollini, D., C. M. Rigsby and E. K. Barto (2012). "Microbes as targets and mediators of allelopathy in plants." Journal of Chemical Ecology **38**(6): 714-727.

Clark, K., I. Karsch-Mizrachi, D. J. Lipman, J. Ostell and E. W. Sayers (2016). "GenBank." Nucleic Acids Research **44**(Database issue): D67-D72.

Clarke, K. and M. Ainsworth (1993). "A method of linking multivariate community structure to environmental variables." Marine Ecology-Progress Series **92**: 205-205.

Clooney, A. G., F. Fouhy, R. D. Sleator, A. O. Driscoll, C. Stanton, P. D. Cotter and M. J. Claesson (2016). "Comparing apples and oranges?: Next generation sequencing and its impact on microbiome analysis." PLoS ONE **11**(2).

Coats, V. C. and M. E. Rumpho (2014). "The rhizosphere microbiota of plant invaders: an overview of recent advances in the microbiomics of invasive plants." Frontiers in microbiology **5**: 368.

Coordinators, N. R. (2013). "Database resources of the national center for biotechnology information." Nucleic Acids Research **41**(Database issue): D8-D20.

Csardi, G. and T. Nepusz (2006). "The igraph software package for complex network research." InterJournal, Complex Systems **1695**(5).

Cui, Q., I. A. Lewis, A. D. Hegeman, M. E. Anderson, J. Li, C. F. Schulte, W. M. Westler, H. R. Eghbalnia, M. R. Sussman and J. L. Markley (2008). "Metabolite identification via the madison metabolomics consortium database." Nature Biotechnology **26**(2): 162-164.

Curtis, T. P., I. M. Head and D. W. Graham (2003). "Peer reviewed: theoretical ecology for engineering biology." Environmental science & technology **37**(3): 64A-70A.

Das, S., P. K. Meher, U. K. Pradhan and A. K. Paul (2017). "Inferring gene regulatory networks using Kendall's tau correlation coefficient and identification of salinity stress responsive genes in rice." Current Science **112**(6): 1257.

Davison, J. (1988). "Plant beneficial bacteria." Nature Biotechnology **6**(3): 282-286.

Defays, D. (1977). "An efficient algorithm for a complete link method." The Computer Journal **20**(4): 364-366.

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**: 10541.

Deng, Y., Z. He, J. D. Van Nostrand and J. Zhou (2008). "Design and analysis of mismatch probes for long oligonucleotide microarrays." BMC Genomics **9**(1): 1.

Deng, Y., Y.-H. Jiang, Y. Yang, Z. He, F. Luo and J. Zhou (2012). "Molecular ecological network analyses." BMC Bioinformatics **13**(1): 113.

Dunne, J. A., R. J. Williams and N. D. Martinez (2002). "Food-web structure and network theory: the role of connectance and size." Proceedings of the National Academy of Sciences **99**(20): 12917-12922.

Eddy, S. R. (1998). "Profile hidden Markov models." Bioinformatics **14**(9): 755-763.

Eddy, S. R. (2011). "Accelerated profile HMM searches." PLoS Computational Biology **7**(10): e1002195.

Eguiluz, V. M., D. R. Chialvo, G. A. Cecchi, M. Baliki and A. V. Apkarian (2005). "Scale-free brain functional networks." Physical review letters **94**(1): 018102.

Ehrenfeld, J. G. (2003). "Effects of exotic plant invasions on soil nutrient cycling processes." Ecosystems **6**(6): 503-523.

Eiler, A., F. Heinrich and S. Bertilsson (2012). "Coherent dynamics and association networks among lake bacterioplankton taxa." The ISME journal **6**(2): 330-342.

el Zahar Haichar, F., C. Marol, O. Berge, J. I. Rangel-Castro, J. I. Prosser, J. m. Balesdent, T. Heulin and W. Achouak (2008). "Plant host habitat and root exudates shape soil bacterial community structure." The ISME journal **2**(12): 1221-1230.

Eppinga, M. B., M. Rietkerk, S. C. Dekker, P. C. De Ruiter and W. H. Van der Putten (2006). "Accumulation of local pathogens: a new hypothesis to explain exotic plant invasions." Oikos **114**(1): 168-176.

Faust, K. and J. Raes (2012). "Microbial interactions: from networks to models." Nature Reviews Microbiology **10**(8): 538-550.

Faust, K., J. F. Sathirapongsasuti, J. Izard, N. Segata, D. Gevers, J. Raes and C. Huttenhower (2012). "Microbial co-occurrence relationships in the human microbiome." PLoS Computational Biology **8**(7): e1002606.

Filzmoser, P. and K. Hron (2009). "Correlation analysis for compositional data." Mathematical Geosciences **41**(8): 905-919.

Finn, R. D., A. Bateman, J. Clements, P. Coggill, R. Y. Eberhardt, S. R. Eddy, A. Heger, K. Hetherington, L. Holm and J. Mistry (2013). "Pfam: the protein families database." Nucleic Acids Research: gkt1223.

Finn, R. D., J. Clements and S. R. Eddy (2011). "HMMER web server: interactive sequence similarity searching." Nucleic Acids Research.

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

Fornell, C. and D. F. Larcker (1981). "Evaluating structural equation models with unobservable variables and measurement error." Journal of marketing research: 39-50.

Frankenberger Jr, W. T. and M. Arshad (1995). Phytohormones in soils: microbial production and function, Marcel Dekker Inc.

Franzosa, E. A., T. Hsu, A. Sirota-Madi, A. Shafquat, G. Abu-Ali, X. C. Morgan and C. Huttenhower (2015). "Sequencing and beyond: integrating molecular 'omics' for microbial community profiling." Nature Reviews Microbiology **13**(6): 360-372.

Friedman, J. and E. J. Alm (2012). "Inferring correlation networks from genomic survey data." PLoS Computational Biology **8**(9): e1002687.

Frostegård, Å., A. Tunlid and E. Bååth (2011). "Use and misuse of PLFA measurements in soils." Soil Biology and Biochemistry **43**(8): 1621-1625.

Fuhrman, J. A. (2009). "Microbial community structure and its functional implications." Nature **459**(7244): 193-199.

Gans, J., M. Wolinsky and J. Dunbar (2005). "Computational improvements reveal great bacterial diversity and high metal toxicity in soil." Science **309**(5739): 1387-1390.

Gao, H., Z. K. Yang, T. J. Gentry, L. Wu, C. W. Schadt and J. Zhou (2007). "Microarray-based analysis of microbial community RNAs by whole-community RNA amplification." Applied and Environmental Microbiology **73**(2): 563-571.

Garbeva, P., J. A. Van Veen and J. D. Van Elsas (2004). "Microbial diversity in soil: selection of microbial populations by plant and soil type and implications for disease suppressiveness." Annu. Rev. Phytopathol. **42**: 243-270.

Gianfreda, L. (2015). "Enzymes of importance to rhizosphere processes." Journal of soil science and plant nutrition **15**: 283-306.

Glass, E. M., J. Wilkening, A. Wilke, D. Antonopoulos and F. Meyer (2010). "Using the metagenomics RAST server (MG-RAST) for analyzing shotgun metagenomes." Cold Spring Harbor Protocols **2010**(1): pdb. prot5368.

Glick, B. R. (1995). "The enhancement of plant growth by free-living bacteria." Canadian Journal of Microbiology **41**(2): 109-117.

Gonzalez, A. and R. Knight (2012). "Advancing analytical algorithms and pipelines for billions of microbial sequences." Current Opinion in Biotechnology **23**(1): 64-71.

Gornish, E. S., N. Fierer and A. Barberán (2016). "Associations between an invasive plant (Taeniatherum caput-medusae, Medusahead) and soil microbial communities." PLoS ONE **11**(9): e0163930.

Grayston, S. J., S. Wang, C. D. Campbell and A. C. Edwards (1998). "Selective influence of plant species on microbial diversity in the rhizosphere." Soil Biology and Biochemistry **30**(3): 369-378.

Guo, X., Y. Zhang, W. Hu, H. Tan and X. Wang (2014). "Inferring nonlinear gene regulatory networks from gene expression data based on distance correlation." PLoS ONE **9**(2): e87446.

Hadwin, A. K. M., L. F. Del Rio, L. J. Pinto, M. Painter, R. Routledge and M. M. Moore (2006). "Microbial communities in wetlands of the Athabasca oil sands: genetic and metabolic characterization." FEMS Microbiology Ecology **55**(1): 68-78.

Hazen, T. C., E. A. Dubinsky, T. Z. DeSantis, G. L. Andersen, Y. M. Piceno, N. Singh, J. K. Jansson, A. Probst, S. E. Borglin and J. L. Fortney (2010). "Deep-sea oil plume enriches indigenous oil-degrading bacteria." Science **330**(6001): 204-208.

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

He, Z., T. J. Gentry, C. W. Schadt, L. Wu, J. Liebich, S. C. Chong, Z. Huang, W. Wu, B. Gu and P. Jardine (2007). "GeoChip: a comprehensive microarray for investigating biogeochemical, ecological and environmental processes." The ISME journal **1**(1): 67-77.

He, Z., L. Wu, X. Li, M. W. Fields and J. Zhou (2005). "Empirical establishment of oligonucleotide probe design criteria." Applied and Environmental Microbiology **71**(7): 3753-3760.

He, Z., M. Xu, Y. Deng, S. Kang, L. Kellogg, L. Wu, J. D. Van Nostrand, S. E. Hobbie, P. B. Reich and J. Zhou (2010). "Metagenomic analysis reveals a marked divergence in

the structure of belowground microbial communities at elevated CO_2." Ecology Letters **13**(5): 564-575.

Hong, S., J. Bunge, C. Leslin, S. Jeon and S. S. Epstein (2009). "Polymerase chain reaction primers miss half of rRNA microbial diversity." The ISME journal **3**(12): 1365-1373.

Hyatt, D., G.-L. Chen, P. F. LoCascio, M. L. Land, F. W. Larimer and L. J. Hauser (2010). "Prodigal: prokaryotic gene recognition and translation initiation site identification." BMC Bioinformatics **11**(1): 1.

Jafarizadeh, M. A., N. Fouladi, H. Sabri and B. R. Maleki (2012). "Investigation of spectral statistics of nuclear systems by maximum likelihood estimation method." Nuclear Physics A **890**: 29-49.

Jeong, H., S. P. Mason, A.-L. Barabási and Z. N. Oltvai (2001). "Lethality and centrality in protein networks." Nature **411**(6833): 41-42.

Kaiser- Bunbury, C. N., S. Muff, J. Memmott, C. B. Müller and A. Caflisch (2010). "The robustness of pollination networks to the loss of species and interactions: a quantitative approach incorporating pollinator behaviour." Ecology Letters **13**(4): 442-452.

Kallmeyer, J., R. Pockalny, R. R. Adhikari, D. C. Smith and S. D'Hondt (2012). "Global distribution of microbial abundance and biomass in subseafloor sediment." Proceedings of the National Academy of Sciences **109**(40): 16213-16216.

Kanehisa, M., M. Araki, S. Goto, M. Hattori, M. Hirakawa, M. Itoh, T. Katayama, S. Kawashima, S. Okuda and T. Tokimatsu (2008). "KEGG for linking genomes to life and the environment." Nucleic Acids Research **36**(suppl 1): D480-D484.

Keane, R. M. and M. J. Crawley (2002). "Exotic plant invasions and the enemy release hypothesis." Trends in Ecology & Evolution **17**(4): 164-170.

Kent, W. J. (2002). "BLAT—the BLAST-like alignment tool." Genome research **12**(4): 656-664.

Kerepesi, C. and V. Grolmusz (2016). "Evaluating the quantitative capabilities of metagenomic analysis software." Current Microbiology **72**(5): 612-616.

Kim, D., A. S. Hahn, S. J. Wu, N. W. Hanson, K. M. Konwar and S. J. Hallam (2015). FragGeneScan-plus for scalable high-throughput short-read open reading frame prediction. 2015 IEEE Conference on Computational Intelligence in Bioinformatics and Computational Biology (CIBCB).

Kinney, J. B. and G. S. Atwal (2014). "Equitability, mutual information, and the maximal information coefficient." Proceedings of the National Academy of Sciences **111**(9): 3354-3359.

Klironomos, J. N. (2002). "Feedback with soil biota contributes to plant rarity and invasiveness in communities." Nature **417**(6884): 67-70.

Kloepper, J. W., J. Leong, M. Teintze and M. N. Schroth (1980). "Enhanced plant growth by siderophores produced by plant growth-promoting rhizobacteria." Nature **286**(5776): 885-886.

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

Kowalski, K. P., C. Bacon, W. Bickford, H. Braun, K. Clay, M. Leduc-Lapierre, E. Lillard, M. K. McCormick, E. Nelson and M. Torres (2015). "Advancing the science of

microbial symbiosis to support invasive species management: a case study on Phragmites in the Great Lakes." Frontiers in Microbiology **6**: 95.

Kurtz, Z. D., C. L. Müller, E. R. Miraldi, D. R. Littman, M. J. Blaser and R. A. Bonneau (2015). "Sparse and compositionally robust inference of microbial ecological networks." PLoS Computational Biology **11**(5): e1004226.

Langille, M. G. I., J. Zaneveld, J. G. Caporaso, D. McDonald, D. Knights, J. A. Reyes, J. C. Clemente, D. E. Burkepile, R. L. Vega Thurber, R. Knight, R. G. Beiko and C. Huttenhower (2013). "Predictive functional profiling of microbial communities using 16S rRNA marker gene sequences." Nature Biotechnology **31**(9): 814-821.

Langmead, B. and S. L. Salzberg (2012). "Fast gapped-read alignment with Bowtie 2." Nature methods **9**(4): 357-359.

Langmead, B., C. Trapnell, M. Pop and S. L. Salzberg (2009). "Ultrafast and memory-efficient alignment of short DNA sequences to the human genome." Genome biology **10**(3): 1.

Lee, Y.-J., J. D. Van Nostrand, Q. Tu, Z. Lu, L. Cheng, T. Yuan, Y. Deng, M. Q. Carter, Z. He and L. Wu (2013). "The PathoChip, a functional gene array for assessing pathogenic properties of diverse microbial communities." The ISME journal **7**(10): 1974-1984.

Levin, S. A. (2006). "Fundamental questions in biology." PLoS Biology **4**(9): e300.

Li, K.-B. (2003). "ClustalW-MPI: ClustalW analysis using distributed and parallel computing." Bioinformatics **19**(12): 1585-1586.

Li, S. and H.-H. Chou (2004). "LUCY2: an interactive DNA sequence quality trimming and vector removal tool." Bioinformatics **20**(16): 2865-2866.

Li, W. and A. Godzik (2006). "Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences." Bioinformatics **22**(13): 1658-1659.

Li, X., Z. He and J. Zhou (2005). "Selection of optimal oligonucleotide probes for microarrays using multiple criteria, global alignment and parameter estimation." Nucleic Acids Research **33**(19): 6114-6123.

Liang, Y., Z. He, L. Wu, Y. Deng, G. Li and J. Zhou (2010). "Development of a common oligonucleotide reference standard for microarray data normalization and comparison across different microbial communities." Applied and Environmental Microbiology **76**(4): 1088-1094.

Liebich, J., C. W. Schadt, S. C. Chong, Z. He, S.-K. Rhee and J. Zhou (2006). "Improvement of oligonucleotide probe design criteria for functional gene microarrays in environmental applications." Applied and Environmental Microbiology **72**(2): 1688-1691.

Lima-Mendez, G., K. Faust, N. Henry, J. Decelle, S. Colin, F. Carcillo, S. Chaffron, J. C. Ignacio-Espinosa, S. Roux and F. Vincent (2015). "Determinants of community structure in the global plankton interactome." Science **348**(6237): 1262073.

Liu, W.-T., T. L. Marsh, H. Cheng and L. J. Forney (1997). "Characterization of microbial diversity by determining terminal restriction fragment length polymorphisms of genes encoding 16S rRNA." Applied and Environmental Microbiology **63**(11): 4516-4522.

Logares, R., S. Sunagawa, G. Salazar, F. M. Cornejo-Castillo, I. Ferrera, H. Sarmento, P. Hingamp, H. Ogata, C. de Vargas, G. Lima-Mendez, J. Raes, J. Poulain, O. Jaillon, P. Wincker, S. Kandels-Lewis, E. Karsenti, P. Bork and S. G. Acinas (2014).

"Metagenomic 16S rDNA Illumina tags are a powerful alternative to amplicon sequencing to explore diversity and structure of microbial communities." Environmental Microbiology **16**(9): 2659-2671.

Loman, N. J., R. V. Misra, T. J. Dallman, C. Constantinidou, S. E. Gharbia, J. Wain and M. J. Pallen (2012). "Performance comparison of benchtop high-throughput sequencing platforms." Nature Biotechnology **30**(5): 434-439.

Lorenzo, P., C. S. Pereira and S. Rodríguez-Echeverría (2013). "Differential impact on soil microbes of allelopathic compounds released by the invasive Acacia dealbata Link." Soil Biology and Biochemistry **57**: 156-163.

Lu, Z., Z. He, V. A. Parisi, S. Kang, Y. Deng, J. D. Van Nostrand, J. R. Masoner, I. M. Cozzarelli, J. M. Suflita and J. Zhou (2012). "GeoChip-based analysis of microbial functional gene diversity in a landfill leachate-contaminated aquifer." Environmental science & technology **46**(11): 5824-5833.

Lugtenberg, B. and F. Kamilova (2009). "Plant-growth-promoting rhizobacteria." Annual Review of Microbiology **63**: 541-556.

Luo, F., Y. Yang, C.-F. Chen, R. Chang, J. Zhou and R. H. Scheuermann (2007). "Modular organization of protein interaction networks." Bioinformatics **23**(2): 207-214.

Luo, F., Y. Yang, J. Zhong, H. Gao, L. Khan, D. K. Thompson and J. Zhou (2007). "Constructing gene co-expression networks and predicting functions of unknown genes by random matrix theory." BMC Bioinformatics **8**: 299.

Luo, F., J. Zhong, Y. Yang, R. H. Scheuermann and J. Zhou (2006). "Application of random matrix theory to biological networks." Physics Letters A **357**(6): 420-423.

Luo, F., J. Zhong, Y. Yang and J. Zhou (2006). "Application of random matrix theory to microarray data for discovering functional gene modules." Physical Review E **73**(3): 031924.

Luyckx, J. and C. Baudouin (2011). "Trehalose: an intriguing disaccharide with potential for medical application in ophthalmology." Clinical Ophthalmology **5**: 577-581.

Mangla, S. and R. M. Callaway (2008). "Exotic invasive plant accumulates native soil pathogens which inhibit native plants." Journal of Ecology **96**(1): 58-67.

Markowitz, V. M., I. M. A. Chen, K. Chu, E. Szeto, K. Palaniappan, M. Pillay, A. Ratner, J. Huang, I. Pagani, S. Tringe, M. Huntemann, K. Billis, N. Varghese, K. Tennessen, K. Mavromatis, A. Pati, N. N. Ivanova and N. C. Kyrpides (2014). "IMG/M 4 version of the integrated metagenome comparative analysis system." Nucleic Acids Research **42**(Database issue): D568-D573.

Maron, J. L., J. Klironomos, L. Waller and R. M. Callaway (2014). "Invasive plants escape from suppressive soil biota at regional scales." Journal of Ecology **102**(1): 19-27.

Massol-Deya, A., R. Weller, L. Rios-Hernandez, J. Zhou, R. F. Hickey and J. M. Tiedje (1997). "Succession and convergence of biofilm communities in fixed-film reactors treating aromatic hydrocarbons in groundwater." Applied and Environmental Microbiology **63**(1): 270-276.

Mates, J. (2000). "Effects of antioxidant enzymes in the molecular control of reactive oxygen species toxicology." Toxicology **153**(1): 83-104.

Maurel, C. (1997). "Aquaporins and water permeability of plant membranes." Annual Review Of Plant Biology **48**(1): 399-429.

McCune, B., J. B. Grace and D. L. Urban (2002). Analysis of ecological communities, MjM software design Gleneden Beach, Oregon.

McLeod, M. L., C. C. Cleveland, Y. Lekberg, J. L. Maron, L. Philippot, D. Bru and R. M. Callaway (2016). "Exotic invasive plants increase productivity, abundance of ammonia- oxidizing bacteria and nitrogen availability in intermountain grasslands." Journal of Ecology **104**(4): 994-1002.

Mehta, M. L. (2004). Random matrices, Academic press.

Mendes, R., M. Kruijt, I. de Bruijn, E. Dekkers, M. van der Voort, J. H. Schneider, Y. M. Piceno, T. Z. DeSantis, G. L. Andersen and P. A. Bakker (2011). "Deciphering the rhizosphere microbiome for disease-suppressive bacteria." Science **332**(6033): 1097-1100.

Metzker, M. L. (2010). "Sequencing technologies—the next generation." Nature Reviews Genetics **11**(1): 31-46.

Mielke, P. W. and K. J. Berry (2007). Permutation methods: a distance function approach, Springer Science & Business Media.

Miethke, M. and M. A. Marahiel (2007). "Siderophore-based iron acquisition and pathogen control." Microbiology and Molecular Biology Reviews : MMBR **71**(3): 413-451.

Muyzer, G., E. C. De Waal and A. G. Uitterlinden (1993). "Profiling of complex microbial populations by denaturing gradient gel electrophoresis analysis of polymerase chain reaction-amplified genes coding for 16S rRNA." Applied and Environmental Microbiology **59**(3): 695-700.

Myung, I. J. (2003). "Tutorial on maximum likelihood estimation." Journal of mathematical Psychology **47**(1): 90-100.

Namiki, T., T. Hachiya, H. Tanaka and Y. Sakakibara (2012). "MetaVelvet: an extension of Velvet assembler to de novo metagenome assembly from short sequence reads." Nucleic Acids Research **40**(20): e155-e155.

Nayfach, S. and K. S. Pollard (2016). "Toward accurate and quantitative comparative metagenomics." Cell **166**(5): 1103-1116.

Neilands, J. B. (1995). "Siderophores: structure and function of microbial iron transport compounds." Journal of Biological Chemistry **270**(45): 26723-26726.

Osborne, D. J. and M. T. McManus (2005). Hormones, signals and target cells in plant development, Cambridge University Press.

Oulas, A., C. Pavloudi, P. Polymenakou, G. A. Pavlopoulos, N. Papanikolaou, G. Kotoulas, C. Arvanitidis and I. Iliopoulos (2015). "Metagenomics: Tools and insights for analyzing next-generation sequencing data derived from biodiversity studies." Bioinformatics and Biology Insights **9**: 75-88.

Overbeek, R., R. Olson, G. D. Pusch, G. J. Olsen, J. J. Davis, T. Disz, R. A. Edwards, S. Gerdes, B. Parrello and M. Shukla (2014). "The SEED and the Rapid Annotation of microbial genomes using Subsystems Technology (RAST)." Nucleic Acids Research **42**(D1): D206-D214.

Pastor-Satorras, R. and A. Vespignani (2001). "Epidemic spreading in scale-free networks." Physical review letters **86**(14): 3200.

Pearson, K. (1901). "LIII. On lines and planes of closest fit to systems of points in space." The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science **2**(11): 559-572.

Peng, Y., H. C. Leung, S.-M. Yiu and F. Y. Chin (2012). "IDBA-UD: a de novo assembler for single-cell and metagenomic sequencing data with highly uneven depth." Bioinformatics **28**(11): 1420-1428.

Peng, Y., H. C. M. Leung, S.-M. Yiu and F. Y. L. Chin (2011). "Meta-IDBA: a de Novo assembler for metagenomic data." Bioinformatics **27**(13): i94-i101.

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

Powell, S., K. Forslund, D. Szklarczyk, K. Trachana, A. Roth, J. Huerta-Cepas, T. Gabaldón, T. Rattei, C. Creevey and M. Kuhn (2013). "eggNOG v4. 0: nested orthology inference across 3686 organisms." Nucleic Acids Research: gkt1253.

Pruitt, K. D., T. Tatusova and D. R. Maglott (2005). "NCBI Reference Sequence (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins." Nucleic Acids Research **33**(Database Issue): D501-D504.

Qiu, X., L. Wu, H. S. Huang, P. E. McDonel, A. V. Palumbo, J. M. Tiedje and J. Zhou (2001). "Evaluation of PCR-generated chimeras: Mutations, and heteroduplexes with 16S rRNA gene-based cloning." Applied and Environmental Microbiology **67**(2): 880-887.

Raaijmakers, J. M. and M. Mazzola (2012). "Diversity and natural functions of antibiotics produced by beneficial and plant pathogenic bacteria." Annual Review of Phytopathology **50**(1): 403-424.

Ram, R. J., N. C. VerBerkmoes, M. P. Thelen, G. W. Tyson, B. J. Baker, R. C. Blake, M. Shah, R. L. Hettich and J. F. Banfield (2005). "Community proteomics of a natural microbial biofilm." Science **308**(5730): 1915-1920.

Rappe, M. S. and S. J. Giovannoni (2003). "The uncultured microbial majority." Annual Reviews in Microbiology **57**(1): 369-394.

Rasuk, M. C., A. B. Fernández, D. Kurth, M. Contreras, F. Novoa, D. Poiré and M. E. Farías (2016). "Bacterial diversity in microbial mats and sediments from the Atacama desert." Microbial Ecology **71**(1): 44-56.

Ravasz, E., A. L. Somera, D. A. Mongru, Z. N. Oltvai and A.-L. Barabási (2002). "Hierarchical organization of modularity in metabolic networks." Science **297**(5586): 1551-1555.

Reimann, C. and P. Filzmoser (2000). "Normal and lognormal data distribution in geochemistry: death of a myth. Consequences for the statistical treatment of geochemical and environmental data." Environmental Geology **39**(9): 1001-1014.

Reinhart, K. O. and R. M. Callaway (2004). "Soil biota facilitate exotic Acer invasions in Europe and North America." Ecological Applications **14**(6): 1737-1745.

Reshef, D. N., Y. A. Reshef, H. K. Finucane, S. R. Grossman, G. McVean, P. J. Turnbaugh, E. S. Lander, M. Mitzenmacher and P. C. Sabeti (2011). "Detecting novel associations in large data sets." Science **334**(6062): 1518-1524.

Reynolds, H. L., A. Packer, J. D. Bever and K. Clay (2003). "Grassroots ecology: plant–microbe–soil interactions as drivers of plant community structure and dynamics." Ecology **84**(9): 2281-2291.

Rhee, S. K., X. Liu, L. Wu, S. C. Chong, X. Wan and J. Zhou (2004). "Detection of genes involved in biodegradation and biotransformation in microbial communities by using 50-mer oligonucleotide microarrays." Applied and Environmental Microbiology **70**(7): 4303-4317.

Rho, M., H. Tang and Y. Ye (2010). "FragGeneScan: predicting genes in short and error-prone reads." Nucleic Acids Research **38**(20): e191-e191.

Rodríguez, H. and R. Fraga (1999). "Phosphate solubilizing bacteria and their role in plant growth promotion." Biotechnology advances **17**(4): 319-339.

Roesch, L. F. W., R. R. Fulthorpe, A. Riva, G. Casella, A. K. M. Hadwin, A. D. Kent, S. H. Daroub, F. A. O. Camargo, W. G. Farmerie and E. W. Triplett (2007). "Pyrosequencing enumerates and contrasts soil microbial diversity." The ISME journal **1**(4): 283-290.

Roh, S. W., G. C. J. Abell, K.-H. Kim, Y.-D. Nam and J.-W. Bae (2010). "Comparing microarrays and next-generation sequencing technologies for microbial ecology research." Trends in biotechnology **28**(6): 291-299.

Rout, M. E. and R. M. Callaway (2009). "An invasive plant paradox." Science **324**(5928): 734-735.

Ruan, Q., D. Dutta, M. S. Schwalbach, J. A. Steele, J. A. Fuhrman and F. Sun (2006). "Local similarity analysis reveals unique associations among marine bacterioplankton species and environmental factors." Bioinformatics **22**(20): 2532-2538.

Sakhr, J. and J. M. Nieminen (2006). "Wigner surmises and the two-dimensional homogeneous Poisson point process." Physical Review E **73**(4): 047202.

Santos, R., D. Hérouart, A. Puppo and D. Touati (2000). "Critical protective role of bacterial superoxide dismutase in Rhizobium–legume symbiosis." Molecular Microbiology **38**(4): 750-759.

Schimel, J. (2016). "Microbial ecology: Linking omics to biogeochemistry." Nature Microbiology **1**: 15028.

Schloss, P. D. and J. Handelsman (2006). "Toward a census of bacteria in soil." PLoS Computational Biology **2**(7): e92.

Schmidt, T. M., E. F. DeLong and N. R. Pace (1991). "Analysis of a marine picoplankton community by 16S rRNA gene cloning and sequencing." Journal of bacteriology **173**(14): 4371-4378.

Scholz, F. W. (1985). "Maximum likelihood estimation." Encyclopedia of Statistical Sciences.

Scholz, M. B., C.-C. Lo and P. S. G. Chain (2012). "Next generation sequencing and bioinformatic bottlenecks: the current state of metagenomic data analysis." Current Opinion in Biotechnology **23**(1): 9-15.

Seshadri, R., S. A. Kravitz, L. Smarr, P. Gilna and M. Frazier (2007). "CAMERA: A community resource for metagenomics." PLoS Biology **5**(3): e75.

Sharpton, T. J., S. J. Riesenfeld, S. W. Kembel, J. Ladau, J. P. O'Dwyer, J. L. Green, J. A. Eisen and K. S. Pollard (2011). "PhylOTU: A high-throughput procedure quantifies microbial community diversity and resolves novel taxa from metagenomic data." PLoS Computational Biology **7**(1): e1001061.

Shi, S., E. Nuccio, D. J. Herman, R. Rijkers, K. Estera, J. Li, U. N. da Rocha, Z. He, J. Pett-Ridge and E. L. Brodie (2015). "Successional trajectories of rhizosphere bacterial communities over consecutive seasons." mbio **6**(4): e00746-00715.

Shi, S., E. E. Nuccio, Z. J. Shi, Z. He, J. Zhou and M. K. Firestone (2016). "The interconnected rhizosphere: High network complexity dominates rhizosphere assemblages." Ecology Letters **19**(8): 926-936.

Shokralla, S., J. L. Spall, J. F. Gibson and M. Hajibabaei (2012). "Next- generation sequencing technologies for environmental DNA research." Molecular ecology **21**(8): 1794-1805.

Silva, G. G. Z., K. T. Green, B. E. Dutilh and R. A. Edwards (2016). "SUPER-FOCUS: a tool for agile functional analysis of shotgun metagenomic data." Bioinformatics **32**(3): 354-361.

Simpson, J. T., K. Wong, S. D. Jackman, J. E. Schein, S. J. M. Jones and I. Birol (2009). "ABySS: a parallel assembler for short read sequence data." Genome research **19**(6): 1117-1123.

Sogin, M. L., H. G. Morrison, J. A. Huber, D. M. 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 **103**(32): 12115-12120.

Song, L., P. Langfelder and S. Horvath (2012). "Comparison of co-expression measures: mutual information, correlation, and model based indices." BMC Bioinformatics **13**(1): 328.

Speed, T. (2011). "A correlation for the 21st century." Science **334**(6062): 1502-1503.

Srivastava, L. M. (2002). Plant growth and development: hormones and environment, Academic Press.

Stacey, G. and N. T. Keen (1995). Plant-microbe interactions, Springer Science & Business Media.

Steele, J. A., P. D. Countway, L. Xia, P. D. Vigil, J. M. Beman, D. Y. Kim, C.-E. T. Chow, R. Sachdeva, A. C. Jones and M. S. Schwalbach (2011). "Marine bacterial, archaeal and protistan association networks reveal ecological linkages." The ISME journal **5**(9): 1414-1425.

Stinson, K. A., S. A. Campbell, J. R. Powell, B. E. Wolfe, R. M. Callaway, G. C. Thelen, S. G. Hallett, D. Prati and J. N. Klironomos (2006). "Invasive plant suppresses the growth of native tree seedlings by disrupting belowground mutualisms." PLoS Biology **4**(5): e140.

Su, X., W. Pan, B. Song, J. Xu and K. Ning (2014). "Parallel-META 2.0: Enhanced metagenomic data analysis with functional annotation, high performance computing and advanced visualization." PLoS ONE **9**(3): e89323.

Suzuki, M. T. and S. J. Giovannoni (1996). "Bias caused by template annealing in the amplification of mixtures of 16S rRNA genes by PCR." Applied and Environmental Microbiology **62**(2): 625-630.

Székely, G. J., M. L. Rizzo and N. K. Bakirov (2007). "Measuring and testing dependence by correlation of distances." The Annals of Statistics **35**(6): 2769-2794.

Szklarczyk, D., A. Franceschini, S. Wyder, K. Forslund, D. Heller, J. Huerta-Cepas, M. Simonovic, A. Roth, A. Santos and K. P. Tsafou (2014). "STRING v10: protein–protein interaction networks, integrated over the tree of life." Nucleic Acids Research: gku1003.

Team, R. C. (2012). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria, 2012, ISBN 3-900051-07-0.

Tiedje, J. M., S. Asuming-Brempong, K. Nüsslein, T. L. Marsh and S. J. Flynn (1999). "Opening the black box of soil microbial diversity." Applied Soil Ecology **13**(2): 109-122.

Tiquia, S. M., L. Wu, S. C. Chong, S. Passovets, D. Xu, Y. Xu and J. Zhou (2004). "Evaluation of 50-mer oligonucleotide arrays for detecting microbial populations in environmental samples." Biotechniques **36**(4): 664-675.

Topçuoğlu, B. D., L. C. Stewart, H. G. Morrison, D. A. Butterfield, J. A. Huber and J. F. Holden (2016). "Hydrogen limitation and syntrophic growth among natural assemblages of thermophilic methanogens at deep-sea hydrothermal vents." Frontiers in Microbiology **7**: 1240.

Tremblay, J., K. Singh, A. Fern, E. S. Kirton, S. M. He, T. Woyke, J. Lee, F. Chen, J. L. Dangl and S. G. Tringe (2015). "Primer and platform effects on 16S rRNA tag sequencing." Frontiers in Microbiology **6**.

Trivedi, P., Z. He, J. D. Van Nostrand, G. Albrigo, J. Zhou and N. Wang (2012). "Huanglongbing alters the structure and functional diversity of microbial communities associated with citrus rhizosphere." The ISME journal **6**(2): 363-383.

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." Molecular Ecology Resources **14**(5): 914-928.

Tu, Q., H. Yu, Z. He, Y. Deng, L. Wu, J. D. Van Nostrand, A. Zhou, J. Voordeckers, Y. J. Lee and Y. Qin (2014). "GeoChip 4: a functional gene- array- based high-throughput environmental technology for microbial community analysis." Molecular ecology resources **14**(5): 914-928.

Turner, T. R., K. Ramakrishnan, J. Walshaw, D. Heavens, M. Alston, D. Swarbreck, A. Osbourn, A. Grant and P. S. Poole (2013). "Comparative metatranscriptomics reveals kingdom level changes in the rhizosphere microbiome of plants." The ISME journal **7**(12): 2248-2258.

Van Der Heijden, M. G., R. D. Bardgett and N. M. Van Straalen (2008). "The unseen majority: soil microbes as drivers of plant diversity and productivity in terrestrial ecosystems." Ecology Letters **11**(3): 296-310.

Van Grunsven, R. H. A., W. H. Van Der Putten, T. Bezemer, W. L. M. Tamis, F. Berendse and E. M. Veenendaal (2007). "Reduced plant–soil feedback of plant species expanding their range as compared to natives." Journal of Ecology **95**(5): 1050-1057.

Van Nostrand, J. D., W.-M. Wu, L. Wu, Y. Deng, J. Carley, S. Carroll, Z. He, B. Gu, J. Luo, C. S. Criddle, D. B. Watson, P. M. Jardine, T. L. Marsh, J. M. Tiedje, T. C. Hazen and J. Zhou (2009). "GeoChip-based analysis of functional microbial communities during the reoxidation of a bioreduced uranium-contaminated aquifer." Environmental Microbiology **11**(10): 2611-2626.

Van Nostrand, J. D., A. Zhou and J. Zhou (2016). "StressChip for monitoring microbial stress response in the environment." Stress and Environmental Regulation of Gene Expression and Adaptation in Bacteria, 2 Volume Set.

Venter, J. C., K. Remington, J. F. Heidelberg, A. L. Halpern, D. Rusch, J. A. Eisen, D. Wu, I. Paulsen, K. E. Nelson and W. Nelson (2004). "Environmental genome shotgun sequencing of the Sargasso Sea." Science **304**(5667): 66-74.

Vieites, J. M., M. E. Guazzaroni, A. Beloqui, P. N. Golyshin and M. Ferrer (2009). "Metagenomics approaches in systems microbiology." FEMS Microbiology Reviews **33**(1): 236-255.

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 **111**(14): 5266-5270.

Waldron, P. J., L. Wu, J. D. V. Nostrand, C. W. Schadt, Z. He, D. B. Watson, P. M. Jardine, A. V. Palumbo, T. C. Hazen and J. Zhou (2009). "Functional gene array-based analysis of microbial community structure in groundwaters with a gradient of contaminant levels." Environmental science & technology **43**(10): 3529-3534.

Wang, F., H. Zhou, J. Meng, X. Peng, L. Jiang, P. Sun, C. Zhang, J. D. Van Nostrand, Y. Deng and Z. He (2009). "GeoChip-based analysis of metabolic diversity of microbial communities at the Juan de Fuca Ridge hydrothermal vent." Proceedings of the National Academy of Sciences **106**(12): 4840-4845.

Weinstock, G. M. (2012). "Genomic approaches to studying the human microbiota." Nature **489**(7415): 250-256.

Whipps, J. M. (2001). "Microbial interactions and biocontrol in the rhizosphere." Journal of experimental Botany **52**(suppl 1): 487-511.

Whitman, W. B., D. C. Coleman and W. J. Wiebe (1998). "Prokaryotes: the unseen majority." Proceedings of the National Academy of Sciences **95**(12): 6578-6583.

Whittaker, R. H. (1960). "Vegetation of the Siskiyou Mountains, Oregon and California." Ecological Monographs **30**(3): 279-338.

Widder, S., K. Besemer, G. A. Singer, S. Ceola, E. Bertuzzo, C. Quince, W. T. Sloan, A. Rinaldo and T. J. Battin (2014). "Fluvial network organization imprints on microbial co-occurrence networks." Proceedings of the National Academy of Sciences **111**(35): 12799-12804.

Wu, L., X. Liu, C. W. Schadt and J. Zhou (2006). "Microarray-based analysis of subnanogram quantities of microbial community DNAs by using whole-community genome amplification." Applied and Environmental Microbiology **72**(7): 4931-4941.

Wu, L., D. K. Thompson, G. Li, R. A. Hurt, J. M. Tiedje and J. Zhou (2001). "Development and evaluation of functional gene arrays for detection of selected genes in the environment." Applied and environmental microbiology **67**(12): 5780-5790.

Wylie, K. M., R. M. Truty, T. J. Sharpton, K. A. Mihindukulasuriya, Y. Zhou, H. Gao, E. Sodergren, G. M. Weinstock and K. S. Pollard (2012). "Novel bacterial taxa in the human microbiome." PLoS ONE **7**(6): e35294.

Xue, K., M. M. Yuan, Z. J. Shi, Y. Qin, Y. Deng, L. Cheng, L. Wu, Z. He, J. D. Van Nostrand, R. Bracho, S. Natali, E. A. G. Schuur, C. Luo, K. T. Konstantinidis, Q. Wang, J. R. Cole, J. M. Tiedje, Y. Luo and J. Zhou (2016). "Tundra soil carbon is vulnerable to rapid microbial decomposition under climate warming." Nature Climate Change **6**(6): 595-600.

Yang, G. Q., W. R. Qiu, Y. N. Jin and F. H. Wan (2013). "Potential allelochemicals from root exudates of invasive Ageratina adenophora." Allelopathy Journal **32**(2): 233.

Yang, J., D. Gong, W. Wang, M. Hu and R. Mao (2012). "Extreme drought event of 2009/2010 over southwestern China." Meteorology and Atmospheric Physics **115**(3): 173-184.

Zhou, A., Z. He, Y. Qin, Z. Lu, Y. Deng, Q. Tu, C. L. Hemme, J. D. Van Nostrand, L. Wu and T. C. Hazen (2013). "StressChip as a high-throughput tool for assessing microbial community responses to environmental stresses." Environmental science & technology **47**(17): 9841-9849.

Zhou, J. (2009). "Predictive microbial ecology." Microbial Biotechnology **2**(2): 154-156.

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

Zhou, J., Y. Deng, F. Luo, Z. He, Q. Tu and X. Zhi (2010). "Functional molecular ecological networks." mBio **1**(4): e00169-00110.

Zhou, J., Y. Deng, F. Luo, Z. He and Y. Yang (2011). "Phylogenetic molecular ecological network of soil microbial communities in response to elevated CO_2." mBio **2**(4): e00122-00111.

Zhou, J., Y. Deng, L. Shen, C. Wen, Q. Yan, D. Ning, Y. Qin, K. Xue, L. Wu and Z. He (2016). "Temperature mediates continental-scale diversity of microbes in forest soils." Nature Communications **7**.

Zhou, J., Y. Deng, P. Zhang, K. Xue, Y. Liang, J. D. Van Nostrand, Y. Yang, Z. He, L. Wu and D. A. Stahl (2014). "Stochasticity, succession, and environmental perturbations in a fluidic ecosystem." Proceedings of the National Academy of Sciences **111**(9): E836-E845.

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): e02288-02214.

Zhou, J., S. Kang, C. W. Schadt and C. T. Garten (2008). "Spatial scaling of functional gene diversity across various microbial taxa." Proceedings of the National Academy of Sciences **105**(22): 7768-7773.

Zhou, J., W. Liu, Y. Deng, Y.-H. Jiang, K. Xue, Z. He, J. D. Van Nostrand, L. Wu, Y. Yang and A. Wang (2013). "Stochastic assembly leads to alternative communities with distinct functions in a bioreactor microbial community." mBio **4**(2): e00584-00512.

Zhou, J., L. Wu, Y. Deng, X. Zhi, Y.-H. Jiang, Q. Tu, J. Xie, J. D. Van Nostrand, Z. He and Y. Yang (2011). "Reproducibility and quantitation of amplicon sequencing-based detection." The ISME journal **5**(8): 1303-1313.

Zhou, J., K. Xue, J. Xie, Y. Deng, L. Wu, X. Cheng, S. Fei, S. Deng, Z. He and J. D. Van Nostrand (2012). "Microbial mediation of carbon-cycle feedbacks to climate warming." Nature Climate Change **2**(2): 106-110.