UNIVERSITY OF OKLAHOMA

GRADUATE COLLEGE

METAGENOMIC INSIGHTS INTO MICROBIAL COMMUNITY RESPONSES TO LONG-TERM ELEVATED CO_2

A DISSERTATION

SUBMITTED TO THE GRADUATE FACULTY

in partial fulfillment of the requirements for the

Degree of

DOCTOR OF PHILOSOPHY

By

QICHAO TU
Norman, Oklahoma
2014

Downloaded from SHAREOK. Persistent Handle URL: https://hdl.handle.net/11244/10457

METAGENOMIC INSIGHTS INTO MICROBIAL COMMUNITY RESPONSES TO LONG-TERM ELEVATED CO_2

A DISSERTATION APPROVED FOR THE
DEPARTMENT OF MICROBIOLOGY AND PLANT BIOLOGY

BY

_____
Dr. Jizhong Zhou, Chair

_____
Dr. Meijun Zhu

_____
Dr. Fengxia (Felicia) Qi

_____
Dr. Michael McInerney

_____
Dr. Bradley Stevenson

© Copyright by QICHAO TU 2014
All Rights Reserved.

Acknowledgements

At this special moment approaching the last stage for this degree, I would like to express my gratitude to all the people who encouraged me and helped me out through the past years.

Dr. Jizhong Zhou, my advisor, is no doubt the most influential and helpful person in pursuing my academic goals. In addition to continuous financial support for the past six years, he is the person who led me into the field of environmental microbiology, from a background of bioinformatics and plant molecular biology. I really appreciated the vast training I received from the many interesting projects I got involved in, without which I would hardly develop my broad experienced background from pure culture microbial genomics to complex metagenomics.

Dr. Zhili He, who played a role as my second advisor, is also the person I would like to thank most. Without his help, I could be still struggling working on those manuscripts lying in my hard drive. I definitely learned a lot from him in organizing massed results into logical scientific work—skills that will benefit me for life.

I also owe great thanks to my committee members—Dr. Meijun Zhu, Dr. Felicia Qi, Dr. Michael McInerney, and Dr. Bradley Steven for serving as my committee members for this long time. As a graduate student trying to develop his own background in a multidisciplinary area, I deeply appreciated their guidance throughout this whole degree period, by providing valuable suggestions for every step in pursuing this degree. The different angles they look at questions greatly benefitted me to analyze problems comprehensively. My special thanks to Dr. Felicia Qi from OU Health

Sciences Center since every time it cost her additional hours to attend my committee meetings driving from and back to OKC.

Also many thanks to many people for their help during my research, especially Dr. Joy Van Nostrand, Dr. James Voordeckers, Dr. Christopher Hemme, Dr. Liyou Wu and Dr. Ye Deng from IEG, Dr. Patrick Chain and Dr. Gary Xie from Los Alamos National Laboratory, Dr. Peter Reich and Dr. Sarah Hobbie from the University of Minnesota. The data analysis methods and scientific writing skills I learned from them would also benefit me for life.

My final sincerest gratitude belongs to my maternal grandparents, my parents and my wife, for their continuous encouragements and support in my whole life. I can imagine the happiness and proudness on his wrinkled face if my grandfather were still alive, when sitting at the riverside in the twilight and talking to his old friends about his young grandson. My wife is the most beautiful, smart and hardworking lady I have ever met. Nothing could compensate the great efforts and sacrifice she has made for this new family.

Table of Contents

Acknowledgements iv

Table of Contents vi

List of Tables xi

List of Figures xii

Abstract xvi

Chapter 1: Introduction 1

1.1 Atmospheric CO_2: the background 1

1.2 Effects of elevated atmospheric CO_2 on macroecosystems 1

1.3 Effects of elevated atmospheric CO_2 on soil microbial communities 3

1.4 Microbial biodiversity and current challenges 6

1.5 Foci of this study 9

Chapter 2: Strain/Species Identification in Metagenomes Using Genome-Specific Markers 13

2.1 Abstract 13

2.2 Introduction 15

2.3 Materials and Methods 18

2.3.1 Data resources 18

2.3.2 Selection of genome-specific markers (GSMs) 19

2.3.3 Specificity evaluation of GSMs 21

2.3.4 Determining the detection limit and true positive thresholds 22

2.3.5 Profiling T2D-/obesity-associated microbial strains 22

2.4 Results 23

2.4.1 Selection of strain/species-specific GSMs 23
2.4.2 Specificity evaluation with mock community metagenomes 26
2.4.3 Specificity evaluation against recently sequenced genomes and body site specific metagenomes 27
2.4.4 Determining the detection limit and true positive calling thresholds for microbial identification using GSMs 30
2.4.5 Comparison with other approaches 31
2.4.6 Metagenomic profiling of type 2 diabete (T2D)-associated microbial strains/species 32
2.4.7 Metagenomic profiling of obesity-associated microbial strains/species ... 35
2.5 Discussion 38
Chapter 3: Long-term Elevated CO_2 Decreases Microbial Biodiversity by Functional Convergence 45
3.1 Abstract 45
3.2 Introduction 46
3.3 Results and discussion 48
3.4 Materials and Methods 59
3.4.1 Site description and sample collection 59
3.4.2 Plant and soil property measurements 60
3.4.3 DNA extraction, purification and quantification 61
3.4.4 Shotgun metagenome sequencing and 16S rRNA gene amplicon sequencing 61
3.4.5 Shotgun data preprocessing, gene prediction and annotation 62

3.4.6 16S amplicon data processing and OTU identification 63
3.4.7 Biodiversity definition and calculation 63
Chapter 4: Fungal Communities Respond to Long-term Elevated CO_2 by Community Reassembly 66
4.1 Abstract 66
4.2 Introduction 68
4.3 Materials and Methods 71
4.3.1 Site description and sample collection 71
4.3.2 DNA extraction, purification and quantification 71
4.3.3 PCR amplification and 454 pyrosequencing 71
4.3.4 Data analysis 72
4.3.5 Co-occurrence ecological network construction and analysis 73
4.3.6 Linking community structure and network topology with soil and plant properties 74
4.4 Results 75
4.4.1 CO_2 effects on soil and plant characteristics 75
4.4.2 Sequence summary 76
4.4.3 Long-term eCO_2 did not change the overall fungal community structure, but increased their diversity 77
4.4.4 The composition of fungal community in grassland soil ecosystems 77
4.4.5 The co-occurrence networks of fungal communities and their responses to eCO_2 80

4.4.6 Linking fungal community structure and network topology with soil and plant properties 83
4.5 Discussion 85
Chapter 5: The Diversity and Co-occurrence Patterns of N_2-fixing Microorganisms in a CO_2 Enriched Grassland Ecosystem 91
5.1 Abstract 91
5.2 Introduction 93
5.3 Materials and Methods 97
5.3.1 Site description and sample collection 97
5.3.2 DNA extraction, purification and quantification 97
5.3.3 PCR amplification and 454 pyrosequencing 97
5.3.4 Data analysis 98
5.3.5 Co-occurrence ecological network construction 99
5.4 Results 100
5.4.1 Effects of eCO_2 on plant biomass, soil N, and nifH gene abundance 100
5.4.2 Sequencing data summary 103
5.4.3 No significant eCO_2 effects on overall nifH-community diversity and structure 103
5.4.4 The taxonomic and phylogenetic composition of nifH genes 105
5.4.5 Co-occurrence ecological networks of nifH communities 107
5.6 Discussion 110
Chapter 6: Summary and Output 117
Appendix A: Supplementary Tables 124

Appendix B: Supplementary Figures 131

References 157

List of Tables

Table 2.1 The list of microbial strains significantly associated with T2D patients with mean normalized hits ≥5 in treatment/control. 34

Table 4.1 Mantel analysis of the relationships between the overall fungal community structure, co-occurrence network topology and individual soil properties. 84

Table 4.2 Mantel test on network connectivity vs. the OTU significances of soil geochemical variables[a] 85

List of Figures

Fig. 2.1 Flowchart of GSM identification processes. First, k-mer database (db) construction. K-mer db representing k-mers that show up in two or more microbial strains and all human genome k-mers was constructed by the *meryl* program. K-mer sizes from 18 to 20 were selected. Second, 50-mer GSMs were generated for selected strains/species. GSMs were then mapped with the k-mer db, and mapped GSMs were filtered. Third, all GSMs were searched against all microbial genomes by BLAST, and GSMs having 85% identity with non-target GSMs were also filtered. 24

Fig. 2.2 Location of the identified GSMs in the genome: (A) strain-specific GSMs; (B) species-specific GSMs. Different colors denote different locations in the genome: blue for GSMs within genes, green for GSMs within intergenic regions, red for GSMs overlapped between a gene and an intergenic region, and purple for unannotated genomes. 25

Fig. 2.3 Specificity and sensitivity evaluation of identified GSMs. (A) Specificity evaluation against recently sequenced genomes. A total of 302 genomes were collected. (B) Specificity evaluation of GSMs targeting microorganisms isolated from different body sites using raw metagenomes reads. GSMs targeting six different body sites (gastrointestinal tract, oral, airways, skin, blood, and urogenital tract) were searched with metagenomes from nine different body sites (stool, subgingival plaque, tongue dorsum, throat, palatine tonsils, anterior nares, left retroauricular crease, right retroauricular crease, and posterior fornix) using MEGABLAST. Numbers denote the percentages of MEGABLAST hits with GSMs targeting each body site. (C) Sensitivity evaluation of GSMs using simulated metagenomes from 695 guts microbial strains.

Simulated metagenomes at seven different coverages (0.01, 0.03, 0.05, 0.1, 0.25, 0.5, and 0.75) were searched against different number of GSMs per strain (1, 5, 10, 25, 50, 100, 200, and 500). The percentages of identified microbial strains were analyzed. 28

Fig. 2.4 Response ratio analysis of obese/lean associated microorganisms at the phylum (A) and strain/species level (B). For strain/species level analysis, only significantly associated ones with normalized hit number ≥ 5 were displayed. * refers to microbial strains that did NOT pass Benjamini-Hochberg false discovery..... 36

Fig. 3.1 CO_2 effects on plant biomass and soil ammonification rate. (A) Stimulated aboveground and root biomass in response to eCO_2 in years 1-4 and years 5-12. (B) Soil ammonification rate between aCO_2 and eCO_2 samples in years 1-4 and years 5-12. Ammonification rate was suppressed by eCO2 in years 1-4, but restored in years 5-12 ($P < 0.1$). 48

Fig. 3.2 Responses of phylogenetic (A-C), taxonomic (D-F), and functional (G-I) diversity to long-term eCO2. For each dimension of biodiversity, all three components, including alpha (local diversity of each sample), beta (dissimilarity among samples), and gamma (regional diversity by pooling samples under same condition) were analyzed. 50

Fig. 3.3 Correlations between soil ammonification rate and aboveground plant biomass (A), functional richness (B), functional diversity (C), phylogenetic diversity (D), taxonomic richness (E), and taxonomic diversity (F). Soil ammonification rate was positively correlated with plant biomass, but negatively with diversity indices ($P<0.1$ except taxonomic diversity). 55

Fig. 3.4 A conceptual framework illustrating how long-term eCO_2 decreases microbial biodiversity. Long-term eCO_2 stimulated plant growth in grassland ecosystems, leading to progressive N limitation in soil. Microorganisms capable to produce NH_4^+ from various sources were favored, resulting in functional convergence of microbial communities. Decreased functional diversity as a result of functional convergence then led to decreased to taxonomic and phylogenetic diversity of microbial communities. Numbers in brackets indicate path coefficients as revealed by path analysis. Bolded numbers were significantly different from zero, based on bootstrap t-test. 57

Fig. 4.1 eCO_2 effects on soil nitrogen (A) and total plant biomass (B). Both soil ammonification rate and plant growth were significantly stimulated after 12 years CO_2 treatment. 76

Fig. 4.2 Response ratio (eCO_2 vs. aCO_2) analysis of fungal OTU changes in response to eCO_2. Only the top 28 most abundant OTUs with relative abundance $\geq$0.3% in aCO_2 or eCO_2 were plotted. Error bar symbols plotted at the right of dashed line indicated increased relative abundances at eCO_2, while error bar symbols plotted at the left of dashed line indicated decreased relative abundances at eCO_2. The genus information as well as actual relative abundance with standard error was also listed..... 80

Fig. 4.3 Community reassembly of sparsely distributed OTUs in the aCO_2 network (A) into highly connected dense modules in the eCO_2 network (B). Colored nodes were the OTUs involved in community reassembly. Teal nodes were the first neighbor of yellow nodes. Different colors refer to different fungal phyla. 82

Fig. 4.4 The sub-network of top 7 most abundant OTUs and their first neighbor nodes in the aCO_2 (A) and eCO_2 (B) networks. Each node represents an OTU, which would be

regarded as a fungal species. The size of nodes represents relative abundance of OTUs. Each link connects two OTUs. Grey links means positive connections, and red means negative connections. Different colors refer to different fungal phyla. The OTUs with top relative abundances were marked with OTU ids..... 83

Fig. 5.1 Effects of long-term eCO_2 on plant biomass (A), soil NO_3^- and NH_4^+ (B), and *nifH* gene abundance (C). Both aboveground and root biomass were averaged from 5 years at the time of sampling, i.e. 2005-2009. Soil NO_3^- and NH_4^+ concentrations were then measured using a semi-open core, one-month *in situ* incubation approach. The abundance of *nifH* genes was obtained from shotgun metagenome datasets by extracting sequences mapped to *nifH* genes. Statistical testing was performed by the Student's t test. ino3 and fno3: initial and final NO_3^- concentration; inh4 and fnh4: initial and final NH_4^+ concentration..... 102

Fig. 5.2 The diversity of *nifH* genes in the grassland ecosystem under ambient CO_2 and elevated CO_2 conditions: (A) Chao1 richness; (B) Shannon evenness; (C) Shannon diversity; (D) phylogenetic diversity. Black line represents the averaged value for each diversity index. Turquoise and light-pink regions represent 95% confidence intervals. 104

Fig. 5.3 Taxonomic (A) and phylogenetic (B) composition of *nifH* genes at both OTU and sequence levels. Taxonomic groups were summarized at the class level..... 107

Fig. 5.4 Co-occurrence modules centered by *nifH* OTUs: (A) *Azospirillum* module; (B) *Mesorhizobium* module; (C) *Bradyrhizobium* module. *nifH* OTUs were represented by diamond shape. 16S rRNA OTUs were represented by circular shape. Different colors refer to different phyla..... 108

Abstract

Understanding how belowground microbial communities respond to increasing atmospheric CO_2 is of crucial importance for global change biology, microbial ecology and predictive biology. However, our understanding of CO_2 effects on microbial communities is still limited, especially due to the immense diversity and as yet-uncultivable nature of most soil microorganisms. By implementing next generation sequencing (NGS) technologies, we comprehensively surveyed the responses of microbial communities to elevated CO_2 (eCO_2) in the BioCON experimental site, a grassland ecosystem, which had been exposed to eCO_2 for 12 years.

In the beginning of this study, it was noticed that computational approaches to identify microbial strains/species from shotgum metagenomes are very limited, thus we have developed a computational algorithm, termed GSMer that identifies genome-specific markers (GSMs) from currently sequenced microbial genomes for strain/species identification in metagenomes. Although GSMer was not very successfully applied in our soil metagenomes due to the extremely low coverage and high diversity of soil microbial communities as well as short sequencing reads from early Illumina GAII performs, it was successfully used to analyze microbial communities with a good coverage of reference genomes, such as human microbiomes. Sensitivity evaluation against synthetic metagenomes with different levels of coverage suggested that 50 GSMs per strain were sufficient to identify most microbial strains with $\geq$ 0.25x coverage, and 10% of selected GSMs in a database should be detected for confident positive callings. Application of GSMs respectively identified 45 and 74 microbial strains/species significantly associated with type-2-diabetes (T2D) patients

and obese/lean individuals from corresponding gastrointestinal tract metagenomes. Our results agreed well with previous studies, but provided strain-level information.

In the following, we analyzed the biodiversity, composition, structure and functional potential of soil microbial communities in response to eCO_2 at multiple (e.g., phylogenetic, taxonomic, genetic, functional) dimensions using next generation sequencing approaches. For each dimension of microbial biodiversity, all components of diversity, including alpha-, beta- and gamma-diversity were analyzed. Our results suggested that long-term eCO_2 decreased the overall microbial biodiversity. Beta-diversity analysis suggested eCO_2 decreased functional beta-diversity, but increased taxonomic and phylogenetic diversity, suggesting long-term eCO_2 selected for microbial function rather than taxonomy. Further meta-analysis suggested that such decreased biodiversity was significantly negatively correlated with increased soil ammonification rate. Moreover, the abundance of gene families involved in ammonium producing pathways increased significantly as well, indicating a functional convergence process as a result of higher demand for biologically available nitrogen (N) by stimulated plant growth as a result of eCO_2. Our findings present evidence that plant-microbe interactions for NH_4^+ as result of progressive nitrogen limitation were an important driving factor, responsible for decreased microbial biodiversity under eCO_2.

We also analyzed the response of fungal communities to long-term eCO_2 by sequencing of 28S rRNA gene amplicons. Long-term eCO_2 did not significantly alter the overall fungal community structure and species richness, but significantly increased community evenness and diversity. Relative abundances of 119 OTUs (~ 27% of the total captured sequences) were changed significantly. More interestingly, significantly

changed OTUs under eCO_2 were associated with increased overall relative abundance of Ascomycota, but decreased relative abundance of Basidiomycota. Co-occurrence ecological network analysis indicated that eCO_2 increased fungal community interactions, as evidenced by higher intermodular and intramodular connectivity and shorter geodesic distance. In contrast, decreased connections for dominant fungal species were observed in the eCO_2 network. Community reassembly of unrelated fungal species into highly connected dense modules was observed. Such changes in the network structure were significantly associated with altered soil and plant properties under eCO_2, especially with increased plant biomass and NH_4^+ availability. This study provides novel insights into our understanding of how eCO_2 shapes soil fungal communities in grassland ecosystems.

Since it was noticed that changes of both belowground microbial biodiversity and fungal communities were significantly correlated with soil ammonification rate, and our previous studies showed that the abundance of *nifH* significantly increased at eCO_2, we hypothesized that N_2-fixing microorganisms would play important roles in response to eCO_2. Therefore, we analyzed N_2-fixing communities by sequencing of *nifH* gene amplicons as well as extraction of *nifH* fragments from shotgun metagenomes. Surprisingly, long-term eCO_2 significantly increased the abundance of *nifH* genes, but did not change the overall *nifH* diversity and diazotrophic community structure. Taxonomic and phylogenetic analysis of amplified *nifH* sequences suggested a high diversity of *nifH* genes in the soil ecosystem, with the majority belonging to cluster I and II *nifH* genes. We then constructed a microbial ecological network using 16S rRNA gene and *nifH* gene profiles. Co-occurrence ecological network analysis suggested a

clear preference of co-occurrence patterns between diazotrophs and other microbial species with different patterns observed for different subgroups of diazotrophs, such as *Azospirillum*/Actinobacteria, *Mesorhizobium*/Conexibacter, and *Bradyrhizobium*/Acidobacteria. This indicated a potential attraction of these non-N_2-fixers by diazotrophs in the soil ecosystem. Interestingly, more complex co-occurrence patterns were found for free-living diazotrophs than commonly known symbiotic diazotrophs, consistent with the physical isolation nature of symbiotic diazotrophs from the environment by root nodules. The study provides novel insights of our understanding microbial ecology of soil diazotrophs in natural ecosystems.

All studies included in this work provided novel insights into the long-term eCO_2 effects on belowground microbial communities, which are of merit for next generation sequencing analysis for microbial ecologists, global change biologists and bioinformaticians.

Keywords: **genome-specific markers; GSMer; strain/species identification; metagenomes; elevated CO_2; microbial biodiversity; functional diversity; functional convergence; progressive nitrogen limitation; fungal community; microbial ecological network; community reassembly; *nifH*; soil diazotrophs; community structure; co-occurrence patterns**

Chapter 1: Introduction

1.1 Atmospheric CO_2: the background

The increasing atmospheric CO_2 concentration has become a serious issue for global climate change. Historically, atmospheric CO_2 concentration was not a critical issue until the first industrial revolution, starting around 1750s, when fossil fuels were extensively combusted as energy sources by industrial plants. By analyzing atmospheric air bubbles trapped in ice cores, atmospheric CO_2 concentrations for past centuries or even up to 160,000 years could be obtained (Raynaud and Barnola 1985; Barnola, Raynaud et al. 1987; Jasper and Hayes 1990; Fischer, Wahlen et al. 1999). Such data provided evidence that the increased CO_2 concentration was due to increased human activities, especially the huge amounts of fossil fuels being burned. According to the latest Intergovernmental Panel on Climate Change (IPCC 2013) report, the atmospheric CO_2 concentration has now reached the highest level in historical records at >400 ppm in year 2013, which is >40% greater than in year 1750 (Stocker 2013). If fossil fuel CO_2 emissions continue at their current rate, the atmospheric CO_2 concentration in year 2100 could increase to as high as 985 ppm (Stocker 2013). Such estimates were consistent even when different predictive models were used. As one of the most important atmospheric components that interact with Earth's biosphere, such an increase in CO_2 concentration would have large impacts on the Earth's ecosystem.

1.2 Effects of elevated atmospheric CO_2 on macroecosystems

Increasing atmospheric CO_2 concentration has accelerated global warming by the greenhouse effect (Solomon 2007) — a process that thermal radiation could be absorbed by greenhouse gases such as CO_2, water vapor, methane and ozone (Kiehl and

Trenberth 1997). As a consequence, macroecosystem communities are greatly affected in the following major aspects: (1) changed phenology and physiology of many organisms; (2) shifts of the range and distribution of species to the poles or higher altitudes; (3) changed composition and interaction within communities; (4) changed structure and dynamics of ecosystems (Walther, Post et al. 2002).

Altered plant communities resulting from eCO_2 are believed to be the major factor that influences belowground microbial communities due to their direct link(s). Responses of plant communities to eCO_2 have mainly been studied by Free Air CO_2 Enrichment (FACE) technologies (Lewin, Hendrey et al. 1994), allowing researchers to measure eCO_2 effects on plant communities in large areas under otherwise natural conditions. Under eCO_2, the growth of plant communities was stimulated (Reich, Knops et al. 2001; Norby, DeLucia et al. 2005; Luo, Hui et al. 2006; Reich, Hobbie et al. 2006; Reich and Hobbie 2013), owing to the increased photosynthesic rates. Both aboveground and belowground plant biomass were stimulated, resulting in increased carbon input to the soil (He, Xu et al. 2010; Xu, He et al. 2013). As a result of stimulated plant growth, biologically available N in soil may decrease, constraining the sustainability of plant community responses to eCO_2 (Hu, Chapin et al. 2001; Luo, Su et al. 2004; Reich, Hobbie et al. 2006; Reich and Hobbie 2013). Although plant growth under eCO_2 could be slower or even decrease due to progressive N limitation, continually increased plant biomass was observed in the BioCON grassland ecosystem after long-term eCO_2 treatment. Since biologically available N in natural soil ecosystems mainly originates from microbial communities, the continually increased

plant biomass suggested an important role that microbial communities may play in maintaining the sustainability of plant community responses to eCO_2.

1.3 Effects of elevated atmospheric CO_2 on soil microbial communities

Rather than being directly affected by increasing atmospheric CO_2, soil microbial communities are more likely affected by other factors caused by eCO_2, such as accumulated carbon (C) input as a result of stimulated plant growth and altered soil physical and chemical properties. As a consequence of changes in plant and soil structure, the microbial biomass, community structure and composition, and community diversity are expected to change.

Decomposition of organic matter into reusable nutrients that are able to enter the various biogeochemical cycles (C, N, S, P), is no doubt the most important role that microbial communities play in the Earth's biosphere. Among these, the C-cycling process is the major pathway by which microbial communities respond to environmental perturbations, including elevated atmospheric CO_2 and warming (Heath, Ayres et al. 2005; Carney, Hungate et al. 2007; Bardgett, Freeman et al. 2008; Drigo, Pijl et al. 2010; He, Xu et al. 2010; Cheng, Booker et al. 2012; Zhou, Xue et al. 2012). Since microbial decomposition is also closely influenced by many other environmental factors, such as pH, soil moisture, aerobic conditions, temperature (Davidson and Janssens 2006), and even litter quality (Ball 1997), contrasting observations of microbial decomposition responses to eCO_2 were also found in different studies. In an experimental scrub-oak ecosystem exposed to doubled CO_2 concentration for six years, stimulated microbial decomposition was found to be responsible for offsetting ≈52% of the additional C that had accumulated at eCO_2 in aboveground and coarse root biomass

with soil C loss being driven by changes in microbial community composition and activities (Carney, Hungate et al. 2007). However, microbial decomposition of organic matters was found to be suppressed in a grassland ecosystem by another study, owing to progressive N limitation in soil caused by stimulated plant growth (Hu, Chapin et al. 2001). These contrasting observations suggested that the response of microbial decomposition to eCO_2 could be ecosystem dependent, and may also be greatly affected by other soil environmental factors.

N cycling consists of important pathways by which microbial communities contribute to ecosystem functioning (Gruber and Galloway 2008). Under eCO_2, plant growth is stimulated and take up more N from the soil, resulting in progressive N limitation (Hu, Chapin et al. 2001; Luo, Su et al. 2004; Reich, Hobbie et al. 2006; Reich and Hobbie 2013). Since microorganisms require various N sources, such as NH_4^+ and NO_3^-, for essential biological activities (Geisseler, Horwath et al. 2010), the changed N availability in soil due to stimulated plant growth is expected to influence belowground microbial communities. Similarly, microbial community activities of N cycling may respond differently to eCO_2. For example, microbial N mineralization from decomposition of organic matters increased under low N availability, while being consistently suppressed by high soil N supply or substrate N concentrations (Craine, Morrow et al. 2007). However, a long-term eCO_2 study suggested contrasting results, that microbial decomposition was suppressed by N limitation in soil (Hu, Chapin et al. 2001).

How elevated atmospheric CO_2 concentration affects soil microbial community composition and structure was not as well studied until recently when advances in high

throughput technologies such as next generation sequencing (NGS) and microbial ecological microarrays made such studies feasible and practical. Studies of microbial community responses to long-term eCO_2 could be generally divided into two major time frames, with the year 2010 as a separator. Prior to 2010, traditional low throughput technologies, such phospholipid fatty acids (PLFA), terminal restriction fragment length polymorphism (T-RFLP), and clone library sequencing of phylogenetic markers, were used to analyze changes of microbial community compositions and structure (Janus, Angeloni et al. 2005; Lipson, Wilson et al. 2005; Carney, Hungate et al. 2007; Chung, Zak et al. 2007; Drigo, Kowalchuk et al. 2008; Lesaulnier, Papamichail et al. 2008; Drigo, Pijl et al. 2010; Feng, Simpson et al. 2010). Since 2010, several high throughput technologies have been developed and applied to analyze the microbial community responses to environmental perturbations. Such technologies include microbial ecological microarrays such as GeoChip (He, Deng et al. 2010; Tu, Yu et al. 2014) and PhyloChip (Brodie, DeSantis et al. 2006; Brodie, DeSantis et al. 2007; Schatz, Phillippy et al. 2010), and 454 and Illumina based next generation sequencing technologies (Mardis 2008; Shendure and Ji 2008; Ansorge 2009; MacLean, Jones et al. 2009; Metzker 2010). As a result, more information regarding how microbial communities respond to eCO_2 has been revealed. Although different results were obtained in various studies, almost all studies suggested a common consensus that the microbial community structure and composition significantly differed between ambient CO_2 and eCO_2 conditions (Janus, Angeloni et al. 2005; Lipson, Wilson et al. 2005; Carney, Hungate et al. 2007; Chung, Zak et al. 2007; Drigo, Kowalchuk et al. 2008; Lesaulnier, Papamichail et al. 2008; Drigo, Pijl et al. 2010; Feng, Simpson et al. 2010; He, Xu et al.

2010; Deng, He et al. 2012; Dunbar, Eichorst et al. 2012; He, Piceno et al. 2012; Xu, He et al. 2013). At the functional gene level, it has been found that abundances of a series of gene families involved in C, N, and phosphorous (P) cycling increased significantly in response to eCO_2 (He, Xu et al. 2010; Xu, He et al. 2013). Specifically, abundances of gene families responsible for labile C degradation, rather than recalcitrant C degradation, increased significantly, suggesting a potential consequence of C sequestration instead of C loss as a result of eCO_2 (He, Xu et al. 2010; Xu, He et al. 2013).

1.4 Microbial biodiversity and current challenges

Biodiversity, both aboveground and belowground, is a major factor responsible for ecosystem multifunctioning and stability (McCann 2000; Hector and Hooper 2002; Tilman, Reich et al. 2006; Hector and Bagchi 2007; Zavaleta, Pasari et al. 2010; Wagg, Bender et al. 2014). Biodiversity in an ecosystem usually encompasses four dimensions: taxonomic, phylogenetic, genetic and functional diversity (Naeem, Duffy et al. 2012). Although more advanced technologies such as next generation sequencing are being developed, characterizing the belowground microbial biodiversity is still extremely challenging, especially for genetic diversity and functional diversity. This is solely due to the extremely diverse and >99% as-yet-uncultivable nature of soil microorganisms in the environment (Rappe and Giovannoni 2003).

Currently, by implementing NGS technologies, the phylogenetic and taxonomic diversity of microbial communities in most ecosystems can be assessed by sequencing of phylogenetic molecular markers, such as 16S rRNA genes. Similarly, the genetic and functional diversity could be analyzed by shotgun metagenome sequencing of whole

community DNA fragments. However, owing to the huge difficulties in data analysis of shotgun metagenomes such as annotation and *de novo* assembly (Scholz, Lo et al. 2012), most current efforts of belowground microbial biodiversity largely focus on phylogenetic and taxonomic diversity. Until recently, a few pilot studies of shotgun soil metagenomes provided some fundamental information regarding the belowground microbial biodiversity. For example, Fierer et al. found that the functional diversity was significantly correlated with the taxonomic diversity in the central US continental (Fierer, Ladau et al. 2013). Similarly, at the global scale, the functional beta-diversity was also significantly correlated with the taxonomic and phylogenetic beta-diversity across multiple biomes (Fierer, Leff et al. 2012). This suggested a mutual deterministic property of functional and taxonomic diversity at the global scale. However, at a local scale, such situation may no longer hold as the dominant microbial species are already determined and microorganisms with specific functions are likely to be selected by the ecosystem. One such example is the soil diazotrophs, which encompass phylogenetically distinct groups of microorganisms, but all of whom possess the same functional trait — the gene *nifH* encoding nitrogenase and functioning as N fixation (Zehr, Jenkins et al. 2003). Thus environmental perturbation at a local scale may behave differently at different dimensions of microbial biodiversity.

Previous studies about eCO_2 effects on microbial biodiversity were mainly carried out by relatively low throughput technologies such as clone library sequencing of 16S rRNA genes, and mainly focused on species richness, leaving other dimensions of microbial biodiversity untapped. For example, no significant changes of microbial diversity were found in two different ecosystems by Lipson et al. and Castro et al.

(Lipson, Wilson et al. 2005; Castro, Classen et al. 2010). However, increased bacterial diversity as a result of eCO_2 was found by Janus et al. and Lesaulnier et al.'s studies (Janus, Angeloni et al. 2005; Lesaulnier, Papamichail et al. 2008), the latter of which also suggested decreased archaeal diversity in the FACE experimental site in Rhinelander, WI, USA (Janus, Angeloni et al. 2005; Lesaulnier, Papamichail et al. 2008). In contrast, the microbial species richness, as measured by 16S rRNA sequencing and PhyloChip, in the BioCON experimental site in Minnesota decreased in response to eCO_2 (Deng, He et al. 2012; He, Piceno et al. 2012). These contrasting results could be due to ecosystem differences, but also experimental procedures. For example, the study by Janus et al. (Janus, Angeloni et al. 2005) was carried out in a forest ecosystem subjected to five years of eCO_2, in which much less plant biomass was returned to the soil than grassland ecosystem and may take longer time for microbial communities to reach a new balance due to decreased N availability in the soil. Similarly, a large amount of decreased N availability was observed in the BioCON site in the first 4 years (see Chapter 2), but restored to a higher level afterwards, suggesting a new balance that microbial communities might have reached to provide more N to stimulated plant growth.

However, the effect of eCO_2 on the functional diversity has not been well studied. According to many recent studies, the functional diversity might be more important for maintaining essential ecosystem functioning. For example, shotgun metagenomic analysis suggested a core functional gene set for human microbiome, but each individual hosts a unique microbial community composition (Turnbaugh, Hamady et al. 2009). The importance of functional diversity in response to environmental

perturbations has not been as well studied for microbial communities as in macro-ecosystems, such as in plant communities. For example, Mokany et al. examined a native grassland ecosystem in Canberra, Australia, and found that functional identity and diversity were more important than community diversity in influencing ecosystem processes (Mokany, Ash et al. 2008). Another study by Suding et al. analyzed >900 species across nine terrestrial ecosystems in North America suggested that functional traits of plants were related with diversity loss due to N fertilization (Suding, Collins et al. 2005). These results indicated that functional diversity could be of crucial importance for microbial community responses to eCO_2, which may also explain varied observations of community changes in the taxonomical composition and structure.

1.5 Foci of this study

Although many studies have been carried out and it is certain that eCO_2 has significant effects on microbial community taxonomic and functional structure and composition, how long-term eCO_2 affects belowground microbial biodiversity is still not clear. In previous studies, the analysis of eCO_2 effects on microbial diversity has mainly focused on microbial species richness, leaving other dimensions and components of biodiversity untapped. A comprehensive analysis of how eCO_2 affects microbial biodiversity should include all four dimensions of biodiversity—phylogenetic diversity, taxonomic diversity, genetic diversity and functional diversity. For each of these dimension of biodiversity, all three components, alpha, beta and gamma, should be considered. In addition to species richness, the diversity and evenness of microbial community should also be analyzed, as they are also important in maintaining ecosystem functioning. Most importantly, the underlying mechanism how eCO_2 affects the microbial biodiversity has

not yet been revealed. By taking advantage of next generation sequencing technologies, this study aimed to reveal the response of soil microbial communities to long-term eCO_2 and their mechanisms in a comprehensive manner. Major results are presented in the following four chapters (2-5).

In Chapter 2, we present a novel *k*-mer based approach, GSMer to identify strain/species–specific markers from currently sequenced microbial genomes, which could be then used for strain/species identification in shotgun metagenomes. Although GSMer was not successfully applied in soil metagenomes due to the extremely low coverage and high diversity of soil microbial communities, it was successfully used to analyze microbial communities with good coverage of reference genomes, e.g., human microbiomes. Sensitivity evaluation against synthetic metagenomes with different coverage suggested that 50 GSMs per strain were sufficient to identify most microbial strains with $\geq$ 0.25x coverage, and 10% of selected GSMs in a database should be detected for confident positive callings. Application of GSMs respectively identified 45 and 74 microbial strains/species significantly associated with type 2 diabetes (T2D) patients and obese/lean individuals from corresponding gastrointestinal tract metagenomes. Our results well agreed with previous studies, but provided more strain-level information.

Chapter 3 presents a comprehensive study about how long-term eCO_2 affects belowground microbial biodiversity. By 16S rRNA amplicon analysis and shotgun metagenome sequencing, we comprehensively analyzed the phylogenetic, taxonomic, genetic and functional diversity of microbial communities in the BioCON experimental site, which had been exposed to long-term eCO_2 treatment for 12 years. The mechanism

of how long-term eCO_2 affects belowground microbial biodiversity was explored. We also proposed a conceptual model to illustrate why the microbial biodiversity decreased as a result of eCO_2.

In Chapter 4, we attempted to analyze the response of the fungal community to long-term eCO_2 in the same BioCON experimental site by sequencing of 28S rRNA amplicons. The diversity, composition, structure, and co-occurrence patterns of belowground fungal communities were comprehensively analyzed. Most interestingly, co-occurrence pattern analysis suggested that fungal communities responded to eCO_2 by community reassembly.

In Chapter 3 and 4, we observed that the changed microbial biodiversity and fungal community were significantly correlated with soil ammonification rate. Therefore, in Chapter 5, we analyzed the response of N_2-fixing microbial communities to eCO_2 by sequencing *nifH* gene amplicons. Together with *nifH* gene fragments extracted from shotgun metagenomes, we found that eCO_2 increased *nifH* gene abundance in soil, though the N_2-fixing microbial community diversity and structure were not significantly changed. We also analyzed co-occurrence patterns between soil diazotrophs and other non-N_2-fixers by combining *nifH* and 16S rRNA gene amplicon sequencing approaches. As a result, we found that free-living soil diazotrophs tend to form more complex co-occurrence networks than symbiotic ones and that different co-occurrence patterns were observed for different diazotrophs, providing novel insights into the microbial ecology of soil diazotrophs in natural ecosystems.

This work comprehensively analyzed the response of microbial communities to long-term eCO_2 and presented novel methods in analyzing next generation sequencing

data. Our results indicated that microbial biodiversity decreases under long-term eCO_2 due to ecosystem functional convergence for ammonium production. Moreover, co-occurrence network analysis suggested that fungal communities respond to eCO_2 by community reassembly. Overall, these results provided novel insights into the long-term eCO_2 effects on belowground microbial communities, and be of merit for next generation sequencing analysis for microbial ecologists, global change biologists and bioinformaticians.

Chapter 2: Strain/Species Identification in Metagenomes Using Genome-Specific Markers

2.1 Abstract

Shotgun metagenome sequencing has become a fast, cheap and high throughput technology for characterizing microbial communities in complex environments and human body sites. However, accurate identification of microorganisms at the strain/species level remains extremely challenging. We present a novel *k*-mer based approach, termed GSMer that identifies genome-specific markers (GSMs) from currently sequenced microbial genomes, which were then used for strain/species level identification in metagenomes. Using 5,390 sequenced microbial genomes, a total of 8,770,321 50-mer strain-specific and 11,736,360 species-specific GSMs were identified for 4,088 strains and 2,005 species (4,933 strains), respectively. The GSMs were first evaluated against mock community metagenomes, recently sequenced genomes, and real metagenomes from different body sites, suggesting that the identified GSMs were specific to their targeting genomes. Sensitivity evaluation against synthetic metagenomes with different coverage suggested that 50 GSMs per strain were sufficient to identify most microbial strains with ≥ 0.25x coverage, and 10% of selected GSMs in a database should be detected for confident positive callings. Application of GSMs respectively identified 45 and 74 microbial strains/species significantly associated with type 2 diabetes (T2D) patients and obese/lean individuals from corresponding gastrointestinal tract metagenomes. Our result agreed well with previous studies, but provided strain-level information. The approach can be directly applied to identify

microbial strains/species from raw metagenomes, without the effort of complex data pre-processing.

Keywords: genome-specific markers; GSMer; strain/species identification; metagenomes

2.2 Introduction

Microorganisms can be found in almost every environment of the Earth's biosphere, and are responsible for numerous biological activities including carbon and nitrogen cycling (Gruber and Galloway 2008), organic contaminant remediation (Robinson, McMullan et al. 2001; Condron, Stark et al. 2010; Chikere, Okpokwasili et al. 2011), and human health and disease. Many human disorders, such as type 2 diabetes, obesity, dental cavities, cancer and some immune-related disease, are known to be related with a single or a group of microorganisms (Turnbaugh, Hamady et al. 2009; Larsen, Vogensen et al. 2010; Ley 2010; Kau, Ahern et al. 2011; Qin, Li et al. 2012; Karlsson, Tremaroli et al. 2013; Schwabe and Jobin 2013). In addition, different strains within the same species may have completely different impacts on human health, such as *E. coli* O157:H7 (Karch, Tarr et al. 2005), which is a highly virulent *E. coli* strain, while most other strains in this same species are non-pathogenic. Thus, characterization and identification of microbial strains/species in the environment and individual human hosts is of crucial importance to reveal human-microbial interactions, especially for patients with microbial mediated disorders. Although different technologies have been developed, the characterization and identification of known microorganisms at strain/species levels remain challenging, mainly due to the lack of high resolution tools and the extremely diverse nature of microbial communities.

Currently, the most commonly used approach to characterize and identify microorganisms in complex environments is to sequence 16S rRNA gene amplicons using universally conserved primers (Wang and Qian 2009). However, due to the high similarity of 16S rRNA gene sequences among different microorganisms, this approach

can only confidently identity microorganisms at high taxonomic levels (e.g., genus, family), but not at the species/strain level, though species identification had been attempted in a few studies with less complex communities (Ravel, Gajer et al. 2010; Conlan, Kong et al. 2012). Even at the genus level, resolution problems with 16S rRNA gene sequences have been reported by many investigators (Janda and Abbott 2007). Therefore, it is necessary to use other molecular markers to identify and characterize microorganisms at the strain/species level in complex environments.

Owing to the advances in next generation sequencing (NGS) technologies, shotgun metagenome sequencing, which tries to capture all DNA/RNA information directly from environmental samples, has been widely applied to characterize microbial communities in various environments (Venter, Remington et al. 2004; Tringe, von Mering et al. 2005; Hemme, Deng et al. 2010; Hess, Sczyrba et al. 2011; Mackelprang, Waldrop et al. 2011), including those of the human body (Turnbaugh, Hamady et al. 2009; Qin, Li et al. 2010; Qin, Li et al. 2012; Karlsson, Tremaroli et al. 2013). Also, with the efforts of the Human Microbiome Project (Peterson, Garges et al. 2009), more than 5,000 sequenced microbial genomes are available as references, making it possible for us to identify and characterize those sequenced microbial strains/species in shotgun metagenomes. However, it is computationally intensive using traditional approaches, such as BLAST (Altschul, Gish et al. 1990) searching or short reads mapping (Hatem, Bozda et al. 2013) metagenomes against currently sequenced microbial genomes (~5,000 genomes), while assembling them into contigs to reduce data sizes is even more challenging (Scholz, Lo et al. 2012). Furthermore, many closely related microbial strains/species share large amounts of genome content, which generates a lot of noise in

assigning short reads to references, resulting in ambiguous observations. In addition, sequencing errors, a common issue in NGS techniques (Victoria Wang, Blades et al. 2012), may also reduce confidence levels and increase ambiguity in assigning reads to reference genome sequences, especially to genomes of highly similar strains. Therefore, there is an urgent need to develop an approach that can accurately identify microbial strains/species from shotgun metagenomes.

Until recently, efforts have been made to unambiguously classify metagenomic reads into species or higher levels using a reduced set of clade-specific genes (Segata, Waldron et al. 2012). However, this approach only incorporates gene coding regions in the genomes, leaving intergenic regions untapped. Moreover, strain level identification of known microorganisms is still not feasible due to the high conservation of clade-specific genes in closely related strains (≥ 94% average nucleotide identity) (Konstantinidis and Tiedje 2005).

In this study, we developed a novel *k*-mer based approach, termed GSMer to identify genome-specific markers from currently sequenced microbial genomes, which could then be used for accurate strain/species-level identification of microorganisms in metagenomes. GSMer firstly identifies a set of GSMs for each genome by rapidly and comprehensively searching all regions in the genome sequence and filtering out non-specific sequences. By searching shotgun metagenomes against these GSMs, the presence/absence and/or the relative abundance of each reference strain/species can be determined. In the following, strain-specific GSMs were evaluated against mock community metagenomes, recently sequenced genomes and real metagenomes from different body sites for specificity. Detection limit and true positive calling thresholds

were also determined. It was then applied to identify microbial strains/species associated with type 2 diabetes and obesity from previously published metagenomes.

2.3 Materials and Methods

2.3.1 Data resources

Reference genome sequences (both finished and draft) targeting 5,390 microbial strains were downloaded from HMP DACC and NCBI GenBank databases. Since human DNA may be the main contamination in human microbiome studies, human genome sequences were also downloaded and included for GSM selection. Duplicated genome sequences from different sources were binned together according to the organism information in GenBank files. Body site information for human associated microbial strains/species was obtained from the HMP DACC project catalogue.

Four mock community metagenomes consisting of 21 bacterial strains were downloaded from NCBI Sequence Read Archive with accession numbers SRR172902, SRR072233, SRR172903 and SRR072232. Among these, two were even mock communities, and two were staggered mock communities. SRA format shotgun metagenomes were converted to FASTA files using the *sra* toolkit. Converted FASTA files were then used to identify microbial strains.

Recently sequenced microbial genomes that were not included in GSM identification were downloaded from the JGI IMG website (http://img.jgi.doe.gov/cgi-bin/w/main.cgi). A total of 302 finished genomes were downloaded. Body-site specific metagenome raw data was downloaded from the HMP DACC website for specificity evaluation of selected GSMs. For each body site, the largest metagenome dataset available was selected. A total of 9 bz2 compressed fastq format metagenomes from

stool (SRS011084, 15.3Gb), subgingival plaque (SRS019029, 1.9Gb), tongue dorsum (SRS011115, 9.9Gb), right retroauricular crease (SRS020263, 4.2Gb), palatine tonsils (SRS019126, 4.2Gb), throat (SRS019127, 2.7Gb), anterior nares (SRS023847, 462.4Mb), left retroauricular crease (SRS017849, 4.7Gb), and posterior fornix (SRS023468, 6.4Gb) were downloaded.

Raw metagenome datasets targeting T2D/control (Qin, Li et al. 2012) gut microbiomes were downloaded from NCBI Sequence Read Archive under accession numbers SRA045646 and SRA050230. Obese/lean metagenome raw data (Turnbaugh, Hamady et al. 2009) were downloaded from NCBI Sequence Read Archive under accession number SRA002775**.** SRA format shotgun metagenomes were converted to FASTA files using the *sra* toolkit. Converted FASTA files were then used to profile disease-associated microbial strains/species.

2.3.2 Selection of genome-specific markers (GSMs)

First, strain-level non-redundant *k*-mers were generated for all collected microbial strains as well as human genomes. *k*-mers occurred in two or more bacterial strains were extracted and combined with all *k*-mers of human genomes as a database for stretch filtering. A *k*-mer table was then built by the *meryl* program adopted from the kmer package (Marçais and Kingsford 2011). To insure high specificity of GSMs and reduce computational cost, *k*-mer sizes ranging from 18 to 20 were used in this study. Second, after transforming GenBank files into fasta files, each reference genome was split into 50-mer fragments without ambiguous nucleotides (such as Ns and other consensus nucleotides). Thus for a genome size of L, the number of 50-mer fragments is as much as L-50. Non-redundant 50-mer fragments were identified and kept for further

filtering. Third, the *k*-mer based approach was used to filter out potentially non-specific 50-mer fragments. One significant feature of non-specific DNA fragments is that they share continuous stretch oligonucleotides with their non-targets. Thus continuous stretch filtering could be used to filter out non-specific 50-mer fragments. Here we employed *k*-mer based strategies for continuous stretch filtering. All *k*-mers in the *k*-mer table were mapped to the 50-mers for each genome by the *mapMers* program (Marçais and Kingsford 2011). Mapped 50-mers were discarded since they shared *k*-mers with other strains. Finally, remaining 50-mers for each genome were then searched against all microbial genomes and human genomes for further global sequence identity filtering using MEGABLAST (Zhang, Schwartz et al. 2000) to search for the closest non-target sequences and recalculate global sequence identities. All 50-mers that share sequence identity $\geq$ 85% between their non-target genomes were discarded. The remaining 50-mers were identified as GSMs.

Species-specific GSMs were identified in a similar way as strain-specific GSMs, but *k*-mer databases were generated at the species level rather than at the strain level. The maximum sequence similarity was calculated between 50-mers and non-target genomes that belong to different species.

To insure each strain/species has enough GSMs from multiple regions for real applications, a minimum of 50 GSMs/strain were desired. For such a purpose, a progressive *k*-mer filtering approach was used. For example, if less than 50 GSMs were identified for a strain at a *k*-mer size of 18, the strain would be subject to GSMs identification using a *k*-mer size of 19 and/or 20. The same procedure was also applied to identify species-specific GSMs. Microbial strains with fewer than 50 GSMs/strain at

both strain/species levels with all three k-mer sizes were excluded for disease-associated strain/species profiling, though more GSMs might be found at longer k-mer sizes.

2.3.3 Specificity evaluation of GSMs

To evaluate the specificity of identified GSMs with known bacterial genomes, GSMs from all available microbial strains/species (50 GSMs/strain) were searched against the mock community consisting of 21 bacterial genomes using MEGABLAST (Zhang, Schwartz et al. 2000). Only perfect matches between metagenome reads and GSMs were considered as effective hits. The same criteria were used for specificity evaluation against recently sequenced microbial genomes.

In order to evaluate the specificity of identified GSMs with unsequenced bacterial genomes, GSMs were separated into different groups by the body site from which the microorganisms had been isolated. Body site information for microbial strains was obtained from the HMP DACC website. Only strains linked to one body site were selected. A total of 6 groups of GSMs were extracted for evaluation, targeting body sites including oral, gastrointestinal tract, airways, skin, urogenital tract, and blood. For each body site, 80 strains with more than 50 GSMs identified were randomly selected. For each randomly selected microbial strain, 50 GSMs were randomly selected, resulting in 24,000 GSMs in total. Metagenomes from different body sites were searched against the selected GSMs using MEGABLAST (Zhang, Schwartz et al. 2000). Only perfect matches between metagenome reads and GSMs were considered as effective hits. It is expected that GSMs targeting microorganisms isolated from one body site will be less likely to be perfectly matched with metagenomes from other

distinct body sites, because different body sites should host different microbial communities.

2.3.4 Determining the detection limit and true positive thresholds

Gut GSMs and their targeted genomes were extracted for evaluation. The identification rates between different numbers of GSMs per strain and different sequencing coverage of microorganisms were analyzed. Simulated metagenomes targeting 695 gut microbial genomes were generated by the Grinder program (Angly, Willner et al. 2012), with coverage ranging from 0.01 to 0.75. Paired-end 100-base reads were randomly generated. Randomly selected of 1, 5, 10, 25, 50, 100, 200, and 500 GSMs per strain were used for evaluation. The simulated metagenomes were searched against GSMs using MEGABLAST (Zhang, Schwartz et al. 2000) for strain/species identification. Only perfect matches were regarded as effective hits.

2.3.5 Profiling T2D-/obesity-associated microbial strains

Raw metagenome reads were downloaded and searched against gastrointestinal tract GSMs using MEGABLAST program (Zhang, Schwartz et al. 2000). Since different microbial strains/species may have different numbers of GSMs, we randomly selected 50 GSMs for each strain for normalization purposes in statistical analysis. Only perfect matches between metagenome reads and GSMs were extracted for statistical analysis. Normalization of blast hits profile representing abundances of microbial strains/species was based on the total number of raw reads, and then further normalized to 10,000,000 (Illumina) or 1,000,000 (454) to avoid too small relative abundance values. Student t-test was applied to evaluate statistical significance of T2D-associated microbial strains/species. Response ratio analysis was used to illustrate obesity-associated

microbial strains/species. Benjamini-Hochberg false discovery rate analysis was applied to detected microbial strains with ≥5 normalized reads to see how many microbial strains remained significant after p-value correction.

2.4 Results

2.4.1 Selection of strain/species-specific GSMs

To our best knowledge, no comparative (meta)genomic tools are currently available to identify genome-specific regions from more than 5,000 microbial genomes. Here we developed a novel approach to identify GSMs of same length by taking advantages of *k*-mer based approaches. Two different criteria, including continuous stretch match length and maximum sequence identity with their non-targets, were used to insure the specificity of GSMs. The whole process of GSMs identification is illustrated in the flowchart (Fig. 2.1). Since the definition of microbial strains and species is still widely debated, strains and species here were defined based on the NCBI classification system, where the binomial nomenclature part defines a species and IDs followed by the binomial name defines a strain.

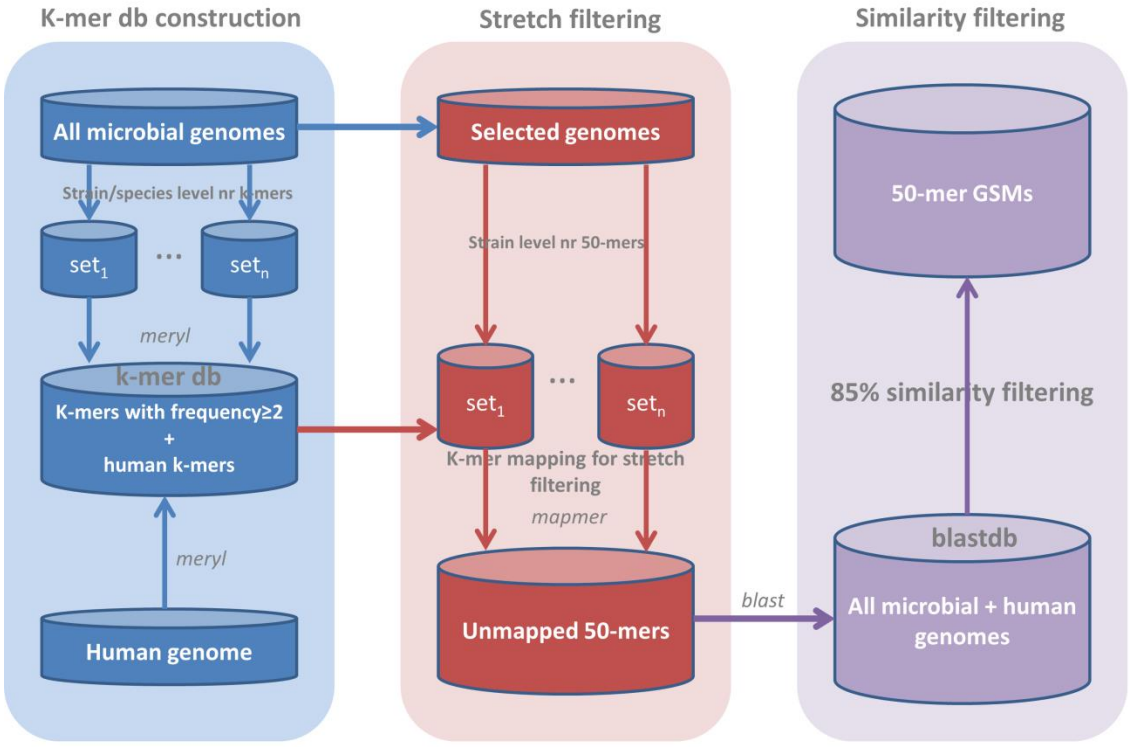


Fig. 2.1 Flowchart of GSM identification processes. First, k-mer database (db) construction. K-mer db representing k-mers that show up in two or more microbial strains and all human genome k-mers was constructed by the *meryl* program. K-mer sizes from 18 to 20 were selected. Second, 50-mer GSMs were generated for selected strains/species. GSMs were then mapped with the k-mer db, and mapped GSMs were filtered. Third, all GSMs were searched against all microbial genomes by BLAST, and GSMs having 85% identity with non-target GSMs were also filtered.

We used two different criteria to identify highly specific GSMs. One is that all GSMs should not have a continuous stretch length ≥ 21-base match with non-target genomes. The other is that all GSMs should not have a sequence identity ≥ 85% with non-target genomes. In order to ensure the identified GSMs are highly specific to their target genomes and reduce the computational time for GSM identification, we started to identify GSMs using a continuous stretch cutoff of 18-mer, then progressively increased the stretch length for genomes without GSMs using previous stretch length, until 20-

mer (Fig. 2.1). The 18-mer starting point was selected for its having relatively large amount (>10 million) of candidate GSMs after *k*-mer continuous stretch filtering, while 17-mer stretch filtering only resulted in ≤20,000 GSMs for ≥ 5000 genomes (Fig. S2.1).

As a result, of the 5,390 microbial strains subject to GSM identification, 4,088 could have ≥ 50 strain-specific GSMs identified. Among them, 2,548 were identified at the 18-mer stretch length, 1,161 at the 19-mer stretch length, and 384 at the 20-mer stretch length. A total of 8,770,321 strain-specific GSMs were identified, among which 6,011,103 (68.5%) were located within genes, 1,657,931 (18.9%) within intergenic regions, 861,008 (9.8%) overlapped between gene and intergenic regions, and 240,092 (2.7%) from unannotated genomes (Fig. 2.2A). Considering the ratio of genes and intergenic regions in a typical bacterial genome (~4.9:1), a higher relative percentage of GSMs were located in or partially in intergenic regions. This also indicated the importance of intergenic regions in bacterial genomes, especially for microbial identification.

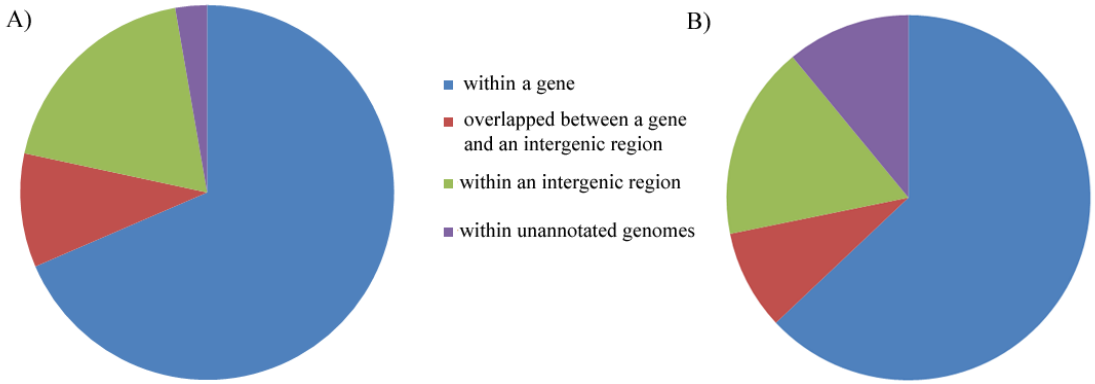


Fig. 2.2 Location of the identified GSMs in the genome: (A) strain-specific GSMs; (B) species-specific GSMs. Different colors denote different locations in the genome: blue for GSMs within genes, green for GSMs within intergenic regions, red for GSMs overlapped between a gene and an intergenic region, and purple for unannotated genomes.

GSMs that target multiple strains in the same species were defined as species-specific GSMs. A total of 11,736,360 GSMs targeting 2,005 species (4,933 strains) were identified. Among them, 1,872 species (3,219 strains) were identified at an 18-mer size, 198 species (1,454 strains) at a 19-mer size, and 48 species (260 strains) at a 20-mer size. About 63% (7,391,847) were located within genes, 8.8% (1,037,718) overlapped between a gene and its intergenic regions, 17.2% (2,016,522) within intergenic regions, and 11% (1,290,273) from unannotated genomes (Fig. 2.2B). This distribution was generally consistent with strain-specific GSMs, suggesting that intergenic regions are important for selecting species-specific GSMs. To select GSMs for the remaining microbial strains without GSMs using the above criteria, modified strategies such as longer stretch length and/or relaxed identity cutoffs could be used.

2.4.2 Specificity evaluation with mock community metagenomes

In order to check the specificity of GSMs with currently sequenced genomes, we first evaluated the selected GSMs against a mock microbial community consisting of 21 bacterial species (Peterson, Garges et al. 2009), of which 16 species had GSMs available. As a result, all 16 (100%) bacterial strains were identified without false positives for the two "even-distributed" mock community data sets sequenced by Illumina and 454 (SRR172902 and SRR072233). Twelve (75%) and fourteen (87.5%) true positives were identified for the two staggered mock community datasets--SRR172903 and SRR072232, respectively. False negative identification in the staggered mock communities was due to the low coverage of these strains (Table S2.1). Three false positives were found in the dataset SRR172903 with only one mapped read for each strain. Of these, two belonged to closely related strains at the same species,

thus it might be caused by the incomplete sequencing of these strains or contamination (Table S2.1). However, these false positives could be effectively removed if a cutoff of identified reads number (e.g., 5) and/or mapped GSM number (e.g., 5) be used.

2.4.3 Specificity evaluation against recently sequenced genomes and body site specific metagenomes

Another question is how specific the GSMs are to unsequenced genomes. This is also critical for true positive callings from metagenomes because the majority of microbial genomes are not yet sequenced, though more than 5,390 microbial strains were used for GSM identification. To evaluate the specificity of GSMs with unsequenced genomes, we collected 302 finished genomes that were recently sequenced (not included in the GSM target strains) and searched them against strain-specific GSMs. A total of 203 (67.2%) genomes were not assigned to any genomes (Fig. 2.3A). Of the 99 (32.8%) genomes assigned to the strains in the GSM database, 75 (24.8%) were assigned to closely related strains in the same species, 14 (4.6%) to the same genus but different species, and only 10 (3.3%) were assigned to different genera (Fig. 2.3A). This suggests that the GSMs identified in this study are even highly specific to unsequenced microbial genomes.

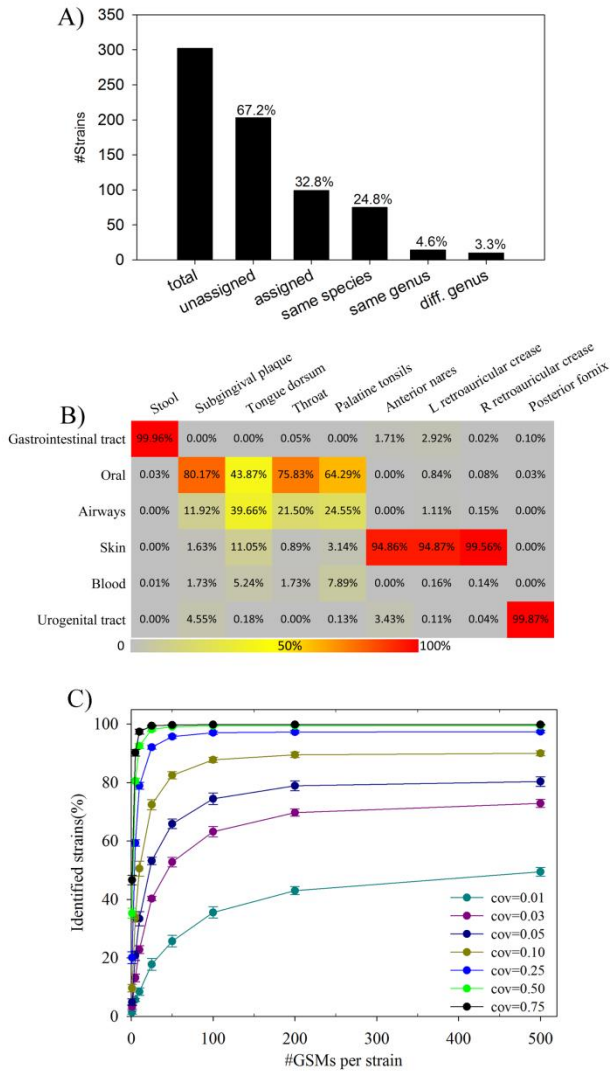


Fig. 2.3 Specificity and sensitivity evaluation of identified GSMs. (A) Specificity evaluation against recently sequenced genomes. A total of 302 genomes were collected. (B) Specificity evaluation of GSMs targeting microorganisms isolated from different body sites using raw metagenomes reads. GSMs targeting six different body sites (gastrointestinal tract, oral, airways, skin, blood, and urogenital tract) were searched with metagenomes from nine different body sites (stool, subgingival plaque, tongue dorsum, throat, palatine tonsils, anterior nares, left retroauricular crease, right retroauricular crease, and posterior fornix) using MEGABLAST. Numbers denote the percentages of MEGABLAST hits with GSMs targeting each body site. (C) Sensitivity evaluation of GSMs using simulated metagenomes from 695 guts microbial strains. Simulated metagenomes at seven different coverages (0.01, 0.03, 0.05, 0.1, 0.25, 0.5, and 0.75) were searched against different number of GSMs per strain (1, 5, 10, 25, 50, 100, 200, and 500). The percentages of identified microbial strains were analyzed.

In addition, we also performed an alternative evaluation to verify the specificity of GSMs, which was less rigorous, but still illustrative. In this test, we hypothesized that microorganisms isolated from one body site would be less likely to be found in another distinctly different body site, based on current studies that different body sites host different microbial communities (Costello, Lauber et al. 2009). Body site information for microbial strains was obtained from the HMP DACC website. GSMs of microbial strains linked with only one of the major six body sites were extracted, though the possibility existed that some strains may also be found in other body sites. Selected GSMs were searched with raw metagenome data from different body sites. It was expected that far fewer hits could be found in other body sites than the particular body site that the selected strains were isolated from. As a result, of the six groups of GSMs targeting different body sites, three are highly specific to their corresponding body sites, and one (blood GSMs) rarely had any hits since it did not have any corresponding metagenomes (Fig. 2.3B). For example, gut GSMs were mainly targeted by stool metagenomes (99.96%); skin GSMs were mainly targeted by metagenomes from anterior nares (94.86%), left retroauricular crease (94.87%), and right retroauricular crease (99.56%); urogenital tract GSMs were mainly targeted by posterior fornix metagenome (99.87%) (Fig. 2.3B). The only exception was that a relatively high number of oral metagenomes, including tongue dorsum (39.66%), throat (21.50%), and palatine tonsils (24.55%) were hit by GSMs targeting microorganisms isolated from airways. Since these oral sites are so closely located and connected with airways, and share similar physiological and functional properties, it is possible for some microorganisms to co-occur in different body sites. This was also evidenced by

previous studies that microbes from oral sites, such as tongue, tonsils, throat, saliva, and gingival plaques, contribute to the colonization in the airways for their important overlap between the upper segments of the digestive and respiratory segments (Segata, Haake et al. 2012). In addition, low BLAST hit numbers were observed between these oral metagenomes and GSMs targeting gut, skin, blood, or urogenital microorganisms, confirming the high possibility that the hits between oral metagenomes and airway GSMs be resulted from their sharing some microorganisms. These results also suggested that the identified GSMs were highly specific to their targeted microorganisms.

2.4.4 Determining the detection limit and true positive calling thresholds for microbial identification using GSMs

Detection limit (sensitivity) is another important issue to identify microbial strains/species from short metagenome sequences in complex environments. There are two major questions associated with sensitivity: (i) At what sequencing coverage the microbial genome could be identified by GSMs? (ii) How many GSMs are required for effective identification of microbial strains/species? In order to answer these two questions, simulated metagenomes with different degrees of genome coverage were generated from sequenced genomes, and then used to determine how many GSMs could be identified. Of the 695 gut microbial genomes subjected to evaluation, about 40% could be detected at 0.01x sequencing coverage level when 100 or more GSMs per strain were used. The value increased to ~90% with 0.1x sequencing coverage and near 100% for 0.25x sequencing coverage with 50 or more GSMs per strain (Fig. 2.3C). And the trend became saturated at 200 GSMs per strain. Overall, our results suggested that

the minimum required GSMs per strain for very low-coverage (≤ 0.25x) sequence data is 100, and 50 for reasonable sequence coverage (≥ 0.25x) sequence data. However, it should be noted reasonable sequencing coverage of a metagenome is necessary for any methods to identify its specific members, especially at the strain/species level.

Another issue related with microbial identification is the threshold for positive calling of an identified strain or species. In order to confidently identify microbial strain/species in a metagenome, a proper threshold of mapped GSMs is necessary. Hence, we examined the distribution of mapped GSM numbers for the simulated metagenomes when 50 and 100 GSMs/strain was used (Fig. S2.2). With 50 GSMs/strain, more than 94% of microbial strains with 0.5x and 0.75x sequencing coverage were identified with 6~50 GSMs (Fig. S2.2 B and C). And with 100 GSMs/strain, more than 95% were identified with 11~100 GSMs (Fig. S2.2 E and F). Even at 0.25x sequencing coverage, more than 75% microbial strains were identified with ≥ 6 GSMs and ≥ 11 GSMs when 50 and 100 GSMs/strain were used, respectively (Fig. S2.2 A and D). In addition, it was found in the specificity evaluation section that ~82% non-specific identifications in metagenomes from different body sites were with ≤ 5GSMs (~2 GSMs/strain). These results suggested that a 10% threshold cutoff (e.g., 5 to 10 GSMs per strain/species) of the number of selected GSMs could be recommended for positive callings. However, in order to detect low coverage microbial strains/species, a lower cutoff could be used with the potential trade-off of increased false positives.

2.4.5 Comparison with other approaches

To our best knowledge, no approaches are yet available to perform strain level analysis of shotgun metagenomes. Here we performed species level analysis for synthetic

shotgun metagenomes generated from the 302 recently sequenced microbial genomes, and compared the results with the current state of the art, MetaPhlAn. Of the 192 microbial species targeted by the synthetic metagenome, MetaPhlAn identified 68 and 69 true positives, and 38 and 41 false positives at sensitive and very sensitive mode, respectively. When ≥ 5 and ≥ 1 mapped GSMs were used as cutoffs, our GSMer approach showed slightly fewer true positives (58 and 62) but much fewer false positives (16 and 21) (Fig. S2.3). Such differences in true positives and false positives should be due to the higher specificity nature of identified GSMs. For both GSMer and MetaPhlAn, about 2/3 of the microbial species targeted by the recently sequenced genomes were not identified, indicating that both GSMer and MetaPhlAn to be a specific tool for sequenced microbial genomes. This also indicated that such limitations in identifying mainly known microbial strains/species could be a common issue for strain/species-level taxonomic identifiers. In order to increase the ability of identifying more microbial strains/species, more newly sequenced microbial genomes need to be included.

2.4.6 Metagenomic profiling of type 2 diabetes (T2D)-associated microbial strains/species

In order to evaluate whether our selected GSMs could be applied to identify disease-associated microbial strains/species in the human body, GSMs targeting human gut microorganisms were searched with raw metagenome data from 345 Chinese individuals with 174 healthy people and 171 diagnosed with T2D (Qin, Li et al. 2012). The previous study with these metagenomes identified 47 T2D-associated MLGs (metagenomic linkage groups), of which 17 were assigned to known bacterial species, 8

to genera, 2 to families, and 1 to order (Qin, Li et al. 2012). Thus it makes us possible to judge the consistency of our results by comparing with this previous study.

With 50 GSMs per strain, 379 microbial strains and 11 species representing 66 strains were found to be present in at least one individual. A total of 45 microbial strains/species was identified to be significantly ($p \leq 0.05$) related with T2D patients, among which 22 had average normalized hits ≥ 5. After Benjamini-Hochberg false discovery rate (FDR) correction, six strains remained to be significantly correlated with T2D (Table 2.1). Since the FDR procedure is closely related with the number of detected microbial strains and all detected microbial strains could be considered independent and uncorrelated, all 22 potential T2D-associated microbial strains with t-test $p \leq 0.05$ without FDR correction were analyzed here. Of them, 14 were enriched in T2D patients, while the remaining 8 were enriched in healthy individuals (Table 2.1). Further literature mining showed that many of the T2D-enriched microbial strains/species were previously identified as potential opportunistic pathogens, such as *Bacteroides caccae* ATCC 43185 (Lozupone, Hamady et al. 2008), *Clostridium bolteae* ATCC BAA-613 (Song, Liu et al. 2004), *Escherichia coli* DEC6E, or not yet well characterized microbial strains that are distinct from currently recognized strains such as those named *Alistipes* sp., *Bacteroides* sp., *Parabacteroides* sp., and *Subdoligranulum* sp.. In addition, the mucin-degrading strain *Akkermansia muciniphila* ATCC BAA-835 was also found to be significantly enriched in T2D patients, which was also observed in the previous study (Qin, Li et al. 2012). In contrast, most microbial strains enriched in healthy individuals belong to butyrate-producing bacteria, such as *Clostridiales bacterium* SS3/4, *Eubacterium rectale* ATCC 33656, *Eubacterium*

rectale DSM 17629, *Faecalibacterium cf. prausnitzii* KLE1255, *Roseburia intestinalis* XB6B4, and *Roseburia inulinivorans* DSM 16841. Two *Prevotella* strains, *Prevotella copri* DSM 18205 and *Prevotella stercorea* DSM 18206, which were reported to be highly associated with carbohydrate consumption (Wu, Chen et al. 2011), were also found to be significantly ($p < 0.05$) enriched in healthy individuals. These results well agreed with previous results based on metagenome-wide association studies (Qin, Li et al. 2012), but provided more detailed information at the strain level.

Table 2.1 The list of microbial strains significantly associated with T2D patients with mean normalized hits ≥5 in treatment/control metagenomes.

Strain	#Mean Mormalized Hits±SDOM		P-value	P-value after FDR correction
	Control	Treatment		
T2D-enriched				
Akkermansia muciniphila ATCC BAA-835	4.79±1.72	18.12±4.58	0.0065	0.07
Alistipes indistinctus YIT 12060	3.60±1.02	8.40±1.84	0.0222	0.15
Alistipes sp. HGB5	3.43±0.40	6.58±1.14	0.0090	0.06
Bacteroides caccae ATCC 43185	29.88±3.39	56.74±8.38	0.0030	0.04
Bacteroides cellulosilyticus DSM 14838	9.27±1.90	17.08±3.31	0.0405	0.17
Bacteroides sp. 2_1_16	2.87±0.57	5.90±1.41	0.0454	0.21
Bacteroides sp. 2_1_33B	4.20±0.55	8.84±2.00	0.0247	0.13
Bacteroides sp. 20_3	15.18±1.96	33.66±4.71	0.0003	0.02
Bacteroides sp. D22	4.03±0.54	6.18±0.78	0.0245	0.15
Clostridium bolteae ATCC BAA-613	3.28±0.53	22.50±9.04	0.0330	0.15
Escherichia coli DEC6E	1.27±0.42	5.47±1.85	0.0261	0.16
Lachnospiraceae bacterium	17.93±2.34	26.61±3.74	0.0492	0.15
Parabacteroides sp. D13	4.51±0.93	8.04±1.05	0.0124	0.06
Subdoligranulum sp. 4_3_54A2FAA	2.04±0.31	6.65±1.49	0.0025	0.05
Control-enriched				
Clostridiales bacterium SS3/4	10.05±0.85	7.35±0.92	0.0318	0.19
Eubacterium rectale ATCC 33656	5.58±1.15	2.91±0.45	0.0319	0.14
Eubacterium rectale DSM 17629	7.07±1.05	3.50±0.52	0.0026	0.05
Faecalibacterium cf. prausnitzii KLE1255	20.46±2.30	12.75±2.03	0.0124	0.14
Prevotella copri DSM 18205	204.12±33.5	106.57±22.4	0.0164	0.11
Prevotella stercorea DSM 18206	58.41±14.61	14.00±5.77	0.0052	0.04
Roseburia intestinalis XB6B4	15.78±2.14	7.30±1.30	0.0008	0.04
Roseburia inulinivorans DSM 16841	34.08±4.61	21.76±3.59	0.0360	0.18

2.4.7 Metagenomic profiling of obesity-associated microbial strains/species

Gut GSMs were then applied to identify obesity-associated microbial strains/species in human gut micorbiomes by searching gut GSMs with metagenomes from 18 individuals, of whom 9 were diagnosed as obese, and the rest were lean/overweight (Turnbaugh, Hamady et al. 2009). The comparison (i.e. obese vs lean/overweight) was carried out in the same manner as the original study (Turnbaugh, Hamady et al. 2009). The previous study found an increased abundance of Actinobacteria and a decreased abundance of Bacteroides in obese individuals, but strain/species level identification of microorganisms associated with obesity was not carried out. Here we intend to identify microbial strains/species associated with obesity, and at the same time to evaluate our results with this previous study by summarizing our data at the phylum level.

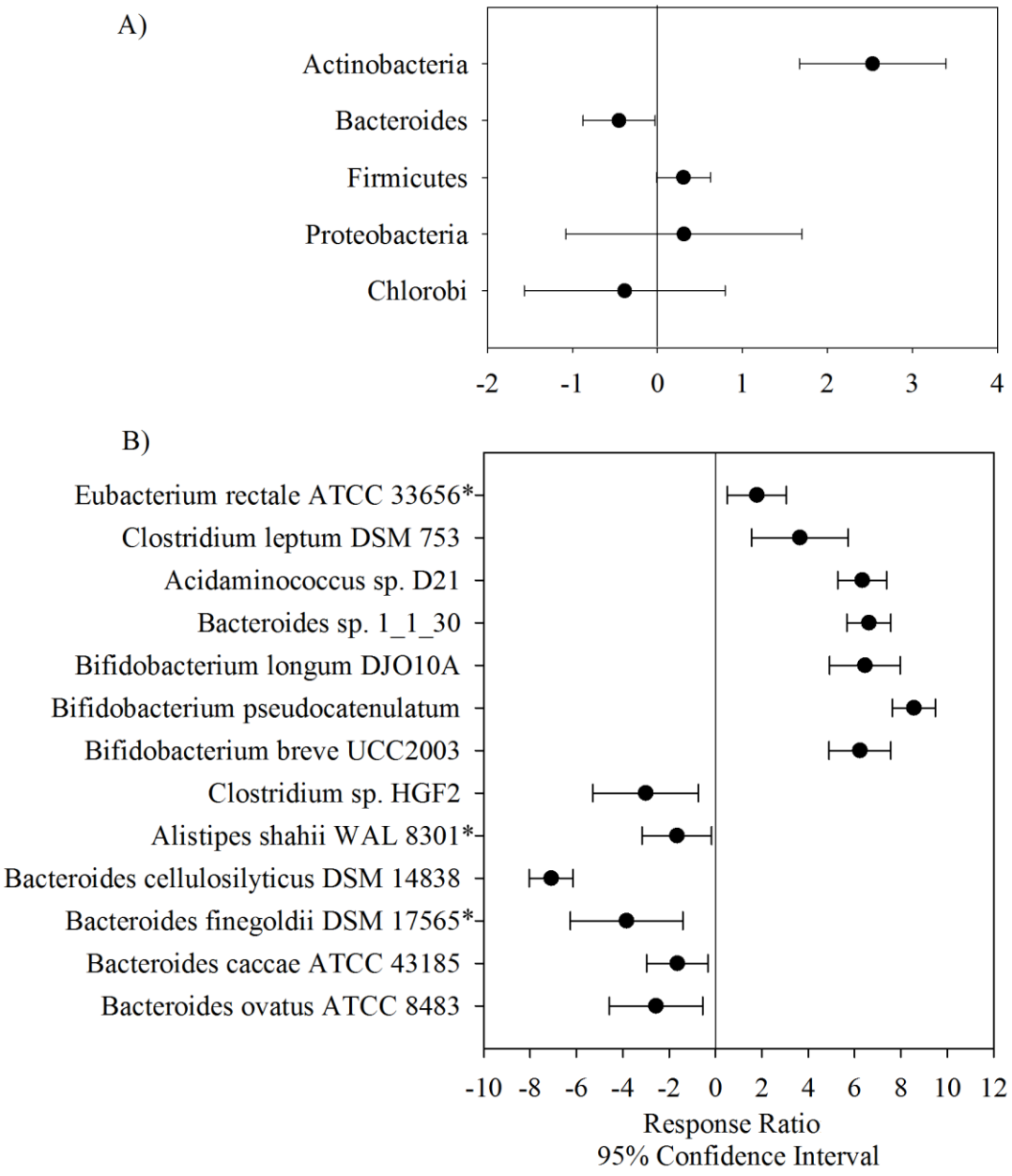


Fig. 2.4 Response ratio analysis of obese/lean associated microorganisms at the phylum (A) and strain/species level (B). For strain/species level analysis, only significantly associated ones with normalized hit number ≥ 5 were displayed. * refers to microbial strains that did NOT pass Benjamini-Hochberg false discovery.

As a result, a total of 159 microbial strains/species were detected in at least one sample in the study. Response ratio analysis showed the relative abundance changes of microbial strains/species between obese and lean/overweight individuals at the 95% confidence interval level. To evaluate whether our results were consistent with the previous one, we first summarized and analyzed the relative abundances of microbial strains/species at the phylum level. A significant lower abundance of Bacteroides and higher abundance of Actinobacteria were found in obese individuals than those in lean/overweight individuals, while no significant changes were observed for microbial phyla such as Firmicutes, Proteobacteria, and Chlorobi between those two groups (Fig. 2.4A). The results were consistent with the previous report using the whole metagenome BLAST searching approach (Turnbaugh, Hamady et al. 2009). We then analyzed the relative abundances of microorganisms at the strain/species level. Relative abundances of a total of 74 strains/species were identified to be significantly ($p < 0.05$) changed in obese/lean individuals. Among these, 13 were found to have an average normalized blast hits number ≥ 5 in obese or lean/overweight individuals. Only three did not pass Benjamini-Hochberg FDR analysis at corrected p-value cutoff of 0.05 (Fig. 2.4B). Of these, six microbial strains were enriched in lean/overweight individuals with five (*Bacteroides cellulosilyticus* DSM 14838, *Bacteroides finegoldii* DSM 17565, *Bacteroides caccae* ATCC 43185, *Bacteroides ovatus* ATCC 8483, and *Alistipes shahii* WAL 8301) in the Bacteroides/Chlorobi group, and one (*Clostridium sp.* HGF2) in Firmicutes. Of the seven obesity-enriched microbial strains/species, three (*Bifidobacterium breve* UCC2003, *Bifidobacterium pseudocatenulatum*, and *Bifidobacterium longum* DJO10A) belonged to Actinobacteria, three (*Acidaminococcus*

sp. D21, *Clostridium leptum* DSM 753, *Eubacterium rectale* ATCC 33656) were Firmicutes, and one (*Bacteroides sp.* 1_1_30) was Bacteroides. Literature search suggests that most microbial strain/species enriched in lean/overweight individuals exhibited potential antibiotic/anti-anaerobic-pathogen resistance abilities (Salyers, Gupta et al. 2004; Wexler 2007), while obesity-enriched microorganisms were mostly probiotics (Schell, Karmirantzou et al. 2002; Sela, Chapman et al. 2008) and butyrate-producing microorganism (Pryde, Duncan et al. 2002; Eckburg, Bik et al. 2005). These results provided new insights for a better understanding of microorganisms associated with obesity at the strain level.

2.5 Discussion

Comparing with other approaches such as BLAST searching against whole genomes for strain/species identification, our approach reduced the searching database to ~0.05% of the whole genomes and minimized noise in strain-level microbial identification. Noise could be introduced when searching metagenomes against whole reference genomes. First, sequencing errors or low quality bases are a common issue in NGS technology (Victoria Wang, Blades et al. 2012). Reduced sequence identity was reported when aligning such error-prone reads against long reference genomes, while the issue can be effectively avoided when searching against short GSMs. Second, the majority of genome content is similar among closely related strains, whereas only small portions are strain/species-specific. Ambiguous assignment of reads to reference genomes in such cases unavoidably introduces great noise for statistical analysis in comparative studies, resulting in ambiguous observations. Since GSMs were extracted from genome-

specific regions, reads not specific to these regions will not be assigned, resulting in more confident microbial identification and statistical analysis.

Specificity is the most important issue for GSM identification. Non-specific GSMs could lead to inaccurate and ambiguous results for strain/species identification in metagenomes. In order to insure highly specific GSMs, several progressive steps were applied. First, more than 5,390 sequenced microbial genomes as well as human genome were used to build *k*-mer databases that feature *k*-mers presenting in two or more genomes, which insures a comprehensive data source in the very beginning. Second, GSMs that can be mapped to any *k*-mers in the databases were discarded, insuring all remaining GSMs do not have any continuous stretch of *k*-mers ($18 \leq k \leq 20$) with non-target genomes. Third, GSMs that share a sequence identity of 85% to their non-target genomes were also discarded, further assuring the specificity of identified GSMs. Fourth, all GSMs were identified as 50-mers, which are shorter than current NGS reads length and can be used for "perfect matching" identification of microbial strains. Finally, our evaluation of GSMs against recently sequenced microbial genomes and metagenomes from different body sites showed they are highly specific.

Sensitivity is another important issue in using GSMs for microbial identification. On one hand, the selected GSMs could be from specific regions of one genome (strain-specific GSMs), or multiple genomes (species-specific GSMs), but not all GSMs could be covered by shotgun metagenome sequences, resulting in false negative detections. On the other hand, most microorganisms in the environment are not sequenced yet, so those incomplete and/or unsequenced genomes may also contain some GSMs identical to those sequenced genomes, leading to false positive detections.

Thus an appropriate number of GSMs and threshold should be determined for confident positive callings of identified microbial strains/species. Our evaluation using simulated metagenomes suggested that a minimum of 50 GSMs per strain and a 10% cutoff for mapped GSMs shall be used for positive callings for most microbial strains at $\geq$ 0.25x sequencing coverage.

A large number of GSMs were identified from intergenic regions for both strain and species-specific GSMs. Intergenic regions comprise about 15% of bacterial genomes (Shabalina, Ogurtsov et al. 2001), and are usually discarded from data analysis in metagenomes at the gene prediction step. The current interest in intergenic regions is focused on exploring novel functional units such as small RNAs, small ORFs, pseudogenes, transposons, integrase binding sites, and repeat elements (Sridhar, Sabarinathan et al. 2011). Our results showed that intergenic regions also contributed heavily to GSMs, suggesting their important roles in identifying microbial strains/species. Thus here we recommended that gene-prediction-free metagenomes should be used for strain/species identification, and the importance of bacterial intergenic regions should be further recognized.

Type 2 diabetes is a complex system level disorder influenced by both genetic and environmental factors (Wellen and Hotamisligil 2005; Scott, Mohlke et al. 2007), as well as the gut microbiome (Musso, Gambino et al. 2011; Qin, Li et al. 2012). Previous studies have suggested significantly different gut microbiome compositions between T2D patients and healthy individuals (Larsen, Vogensen et al. 2010), as well as a group of microbes significantly associated with T2D patients (Qin, Li et al. 2012). By searching metagenome raw reads against 34,750 randomly selected GSMs targeting 695

gut microbial strains, we identified 390 microbial strains/species present in at least one individual. The 45 microbial strains/species significantly associated with T2D were highly consistent with the previous metagenome-wide association study, showing that more "bad" microbes were enriched in T2D patients while more "good" microbes were enriched in healthy individuals. Comparing to the MLG approach, one shortage of our approach is that only sequenced microbial strains/species can be identified, but disease-associated markers from unknown species are not targeted. However, this problem can be solved as more reference genomes are sequenced.

Obesity is a genetically, environmentally and microbially associated energy imbalance disorder in the human body. Studies implementing 16S rRNA sequencing as well as shotgun metagenomes demonstrated significant links between the relative abundances of Actinobacteria, Bacteroides, Firmicutes and obese hosts (Wexler 2007; Turnbaugh, Hamady et al. 2009), such as increased Actinobacteria abundance and Firmicutes/Bacteroides ratio, and decreased Bacteroides. Our phylum level analysis of identified microbial strains using GSMs well agreed with these findings. Intriguingly, unlike the increased opportunistic pathogenic microbes in T2D patients, our strain level analysis showed obese and lean/overweight individuals were associated with different groups of "good" microbes: higher probiotics and butyrate-producing bacteria in obese individuals for maintaining a healthy gut microbiome (Collado, Isolauri et al. 2009) and providing energy source for intestinal epithelial cells (Hamer, Jonkers et al. 2008), and higher antibiotic/anti-anaerobic-pathogen bacteria in lean/overweight individuals. These observations suggested that both obese and lean/overweight individuals host a healthy

gut microbiota, but were enriched by different groups of microbes that harbor different functions.

Both species- and strain-specific GSMs were provided in this study, for the purposes of microbial species and strain identification in metagenomes. Since the majority of currently sequenced microbial strains were covered by the identified GSMs, we expect the method could also be applied to analyze metagenomes from other environments, with the aim to identify sequenced microbial strains/species. However, potential problems may exist, especially for complex microbial communities from environments with limited coverage of reference genomes such as soil, for which the majority of microbial strains are still not yet cultivated and most microbial strains are sequenced with low coverage owing to the extremely high diversity of the community. Such problems would lead to higher false positives and low number of confidently identified microbial strains/species. Thus we recommend mainly using the developed GSMer approach for metagenomes with good coverage of reference genomes such as human microbiome. For complex metagenomes without good coverage of reference genomes, high level taxonomic classifiers (e.g., MEGAN (Huson, Mitra et al. 2011)) should be used for comprehensive data analysis, while high resolution identifiers like GSMer can be used to identify known microbial strains/species with $\geq$ 0.25x coverage. Even with species-specific GSMs, it seems that the majority of novel microbial strains/species still cannot be identified by such high resolution taxonomic identifiers, which is also the same case for MetaPhlAn, though some of them could be assigned to their nearest neighbors. With more novel microbial species/strains being isolated and

sequenced, we expect that such problems could be effectively solved by incorporating more novel microbial genomes.

In conclusion, the GSMer approach we developed here can be used for direct, rapid and accurate identification of microorganisms at the strain/species level from metagenomes, providing a general tool for analysis of metagenome sequencing data. This approach does not require any efforts for preprocessing of huge deluges of reads, including quality trimming, gene prediction, metagenome assembly, and protein-domain matching. In addition, with the advantage of directly taking raw reads, it has the potential to detect microbial strains/species present in low abundances, which are hardly assembled. Although only 50-mer GSMs with very strict parameters were identified and evaluated here, longer GSMs are also supported by the approach with more relaxed parameters. In addition, both gene and intergenic regions were used for GSM selection, expanding the detection ability of microbial strain/species. With more reference genomes being sequenced owing to the progress of HMP project (Human Microbiome Project Consortium 2012; Human Microbiome Project Consortium 2012), strain/species level identification of microorganisms is highly demanded, such as clinical diagnosis for patients with microbial related disorders. Our approach provides a great potential in solving such problems. By integrating such small database with NGS sequencing platforms, instant detection of microbial strains/species is also possible. When applied properly, the method can also be used to select probes for microbial ecological microarrays, which also faces great challenges with huge amount of sequences available.

Availability

All source code for GSMer and testing datasets as well as identified strain/species-specific GSMs could be found at https://github.com/qichao1984/GSMer and http://ieg.ou.edu/GSMer. A semiannual update to cover more newly sequenced genomes is projected. A full list of 50-mer strain/species-specific GSMs identified for all microbial strains can also be downloaded at the above website.

Chapter 3: Long-term Elevated CO_2 Decreases Microbial Biodiversity by Functional Convergence

3.1 Abstract

The belowground microbial biodiversity determines ecosystem multifunctioning (Fierer, Strickland et al. 2009; Wagg, Bender et al. 2014). However, how microbial biodiversity is affected by increasing atmospheric CO_2 remains largely unknown. Through metagenomic analysis of microbial communities in an experimental grassland ecosystem that had been exposed to elevated CO_2 (eCO_2) treatment for 12 years, we found that long-term eCO_2 decreased microbial biodiversity by functional convergence, rather than taxonomy. Such decreased microbial biodiversity was significantly correlated with enhanced soil NH_4^+ as a result of stimulated plant growth. Our findings present evidence that plant-microbe interactions for NH_4^+, as result of progressive nitrogen limitation, are an important driving factor responsible for decreased microbial biodiversity under eCO_2.

Keywords: elevated CO_2; microbial biodiversity; functional diversity; functional convergence; progressive nitrogen limitation

3.2 Introduction

The global atmospheric CO_2 concentration has increased by at least 40% since the industrial revolution began and is likely to increase further due to fossil fuel combustion and land-use changes (Stocker 2013). This increased CO_2 concentration has already significantly affected the Earth's ecosystem, by increasing the Earth's temperature (Stocker 2013), stimulating plant growth (Reich, Knops et al. 2001; Norby, DeLucia et al. 2005; Luo, Hui et al. 2006; Reich and Hobbie 2013), and changing belowground microbial communities (Janus, Angeloni et al. 2005; Lipson, Wilson et al. 2005; Lipson, Blair et al. 2006; Carney, Hungate et al. 2007; Chung, Zak et al. 2007; Bardgett, Freeman et al. 2008; Drigo, Kowalchuk et al. 2008; Lesaulnier, Papamichail et al. 2008; Castro, Classen et al. 2010; Drigo, Pijl et al. 2010; Feng, Simpson et al. 2010; He, Xu et al. 2010; Cheng, Booker et al. 2012; Dunbar, Eichorst et al. 2012; Eisenhauer, Cesarz et al. 2012). A general consensus has been reached by previous studies that microbial communities mainly respond to climate change through C-cycling processes (Heath, Ayres et al. 2005; Carney, Hungate et al. 2007; Bardgett, Freeman et al. 2008; Drigo, Pijl et al. 2010; He, Xu et al. 2010; Cheng, Booker et al. 2012; Zhou, Xue et al. 2012). However, the way in which microbial biodiversity changes as a result of eCO_2 remains uncertain, as contrasting studies have shown increased, decreased or no change in microbial diversity under eCO_2 regimes (Janus, Angeloni et al. 2005; Lipson, Wilson et al. 2005; Lipson, Blair et al. 2006; Lesaulnier, Papamichail et al. 2008; Castro, Classen et al. 2010; Dunbar, Eichorst et al. 2012; Eisenhauer, Cesarz et al. 2012; He, Piceno et al. 2012). This is a critical issue because the belowground microbial biodiversity is intimately linked to aboveground biodiversity (Wardle, Bardgett et al. 2004) and

determines ecosystem multifunctioning (Fierer, Strickland et al. 2009; Wagg, Bender et al. 2014).

Under eCO_2, two major processes—accumulated C input (Reich, Knops et al. 2001; Norby, DeLucia et al. 2005; Luo, Hui et al. 2006; Reich and Hobbie 2013) and progressive N limitation (Hu, Chapin et al. 2001; Norby and Luo 2004; Reich, Hobbie et al. 2006; Norby, Warren et al. 2010; Reich and Hobbie 2013) resulting from stimulated plant growth, may greatly affect belowground microbial communities in a controversial manner. In the first process, increased C input into soil provides more energy source and stimulates microbial activities (Heath, Ayres et al. 2005; Carney, Hungate et al. 2007), thus reducing competition and increasing overall microbial biodiversity. In the second process, enhanced N uptake by stimulated plant growth could lead to progressive N limitation in natural ecosystems (Hu, Chapin et al. 2001; Norby and Luo 2004; Reich, Hobbie et al. 2006; Norby, Warren et al. 2010; Reich and Hobbie 2013), resulting in increased plant-microbe competition for N and suppressed microbial activities (Hu, Chapin et al. 2001), thus decreasing the overall microbial biodiversity. Previous observations (Janus, Angeloni et al. 2005; Lipson, Wilson et al. 2005; Lipson, Blair et al. 2006; Lesaulnier, Papamichail et al. 2008; Castro, Classen et al. 2010; Dunbar, Eichorst et al. 2012; Eisenhauer, Cesarz et al. 2012) of changed microbial biodiversity under eCO_2, mainly focused on species abundance and richness based on small datasets which may be inadequate to draw conclusions for the immense belowground microbial community. Most importantly, other dimensions of microbial biodiversity (e.g., genetic, functional) and the underlying mechanisms responsible for changed biodiversity have not well explored.

3.3 Results and discussion

BioCON (Biodiversity, CO2 and Nitrogen) is a well-established long-term experiment to study ecosystem responses to global climate changes (Reich, Knops et al. 2001). Over ten years of elevated CO_2 treatment (elevated by 180 μmol/mol), plant growth was continuously stimulated (Fig. 3.1A) (Reich, Tilman et al. 2012; Reich and Hobbie 2013), an effect sustained by increased ammonification rate in soil (Fig. 3.1B) (Reich and Hobbie 2013). Interestingly, both nitrification and ammonification rates (N mineralization) were lower under eCO_2 during the first four years ($P<0.1$, Fig. 3.1B, Fig. S3.1AB), confirming suppressed microbial activities by low N availability (Hu, Chapin et al. 2001). More interestingly, nitrification and ammonification rates recovered to a higher level after four years, suggesting restored microbial ammonification and decomposition activities by accumulated carbon input (Reich and Hobbie 2013). This indicated a critical role that soil microbial communities play in combating nitrogen limitation constraints (Norby and Luo 2004; Reich, Hobbie et al. 2006; Reich and Hobbie 2013) by stimulated plant growth in response to long-term eCO_2.

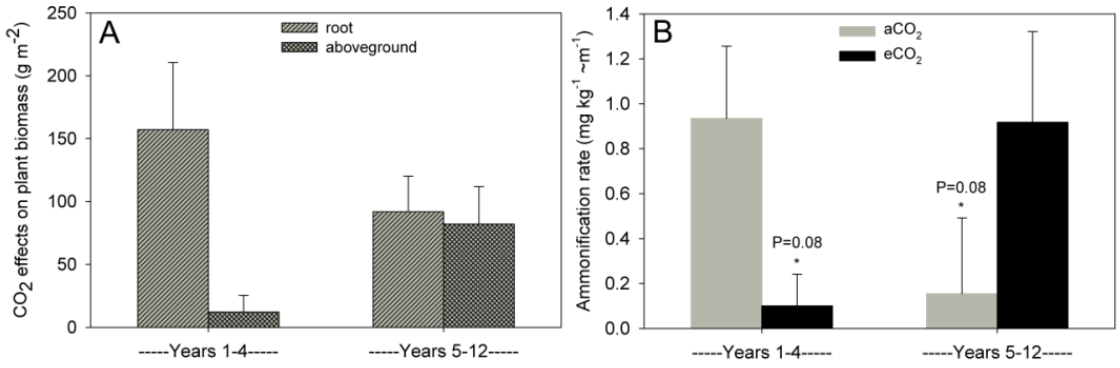


Fig. 3.1 CO_2 effects on plant biomass and soil ammonification rate. (A) Stimulated aboveground and root biomass in response to eCO_2 in years 1-4 and years 5-12. (B) Soil ammonification rate between aCO_2 and eCO_2 samples in years 1-4 and years 5-12.

Ammonification rate was suppressed by eCO2 in years 1-4, but restored in years 5-12 ($P < 0.1$).

In order to examine how long-term eCO_2 affects soil microbial biodiversity, 24 soil samples (0-15cm, 12 from aCO_2, 12 from eCO_2, all with 16 plant species and under ambient N) were collected from the BioCON experimental site. By implementing NGS technologies of sequencing 16S rRNA genes and shotgun metagenomes, we comprehensively surveyed the responses of microbial biodiversity, including phylogenetic, genetic, taxonomic and functional diversity, to long-term eCO_2 in this experimental grassland ecosystem. For each dimension of biodiversity, all three components, including alpha-diversity (local diversity of each sample), beta-diversity (dissimilarity among samples), and gamma-diversity (regional diversity by pooling samples under same condition) were analyzed.

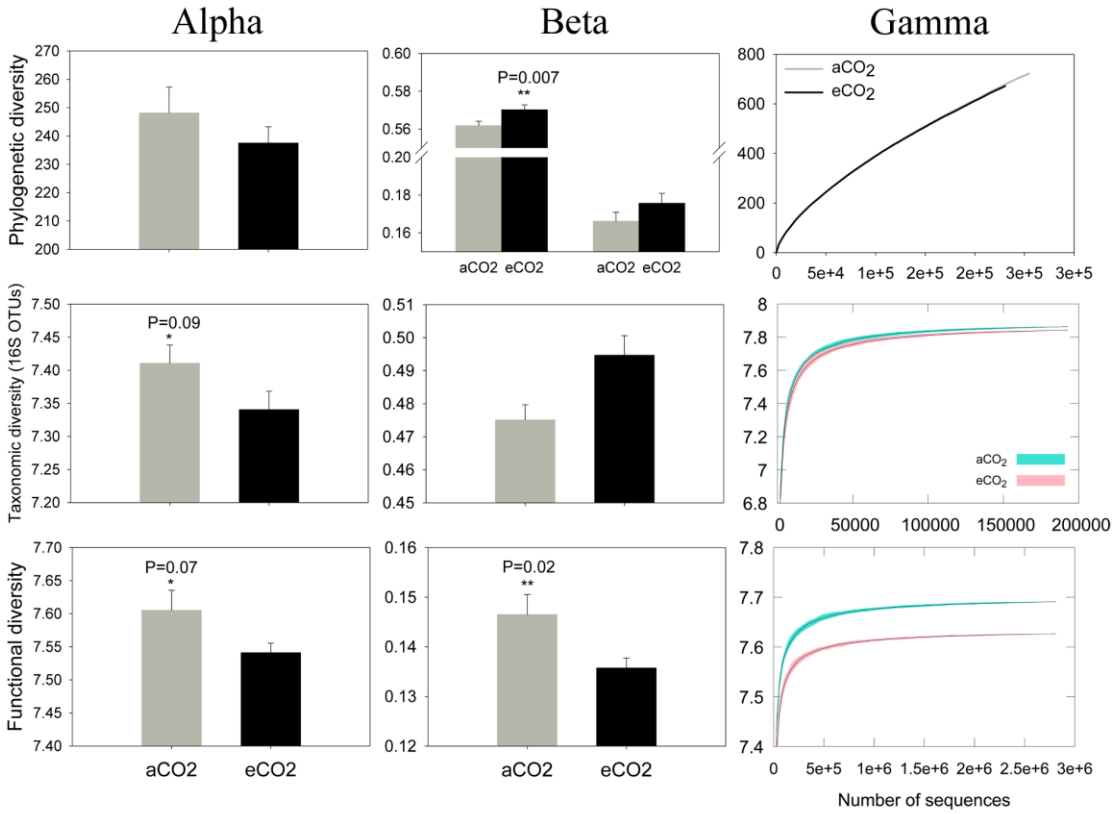


Fig. 3.2 Responses of phylogenetic (A-C), taxonomic (D-F), and functional (G-I) diversity to long-term eCO2. For each dimension of biodiversity, all three components, including alpha (local diversity of each sample), beta (dissimilarity among samples), and gamma (regional diversity by pooling samples under same condition) were analyzed.

Phylogenetic diversity. The microbial community phylogenetic diversity was assessed by 16S rRNA amplicon sequencing. A total of 16,633 non-chimeric OTUs were obtained and subjected to phylogenetic diversity analysis (Fig. 3.2 A-C). Although it was not significant, higher phylogenetic alpha-diversity was observed in aCO_2 samples than that in eCO_2 samples with the same amount of randomly selected sequences for analysis. Unweighted UniFrac distance analysis suggested significantly higher beta-diversity among eCO_2 samples than that among aCO_2 samples. However, significance was eliminated by weighted UniFrac distance analysis, suggesting that rare species

could be an important factor responsible for the changes of phylogenetic diversity. Owing to the opposite trends of alpha- and beta-diversity, no obvious changes of phylogenetic gamma-diversity was found between aCO_2 and eCO_2 samples with any amount of randomly sampled sequences.

Genetic diversity. Genetic diversity was measured by shotgun metagenome sequencing with genetic groups defined as predicted genes sharing ≥90% sequence identity over a ≥30% aligned region. Despite insignificance, each aCO_2 sample encompassed an average of 300,000 more genetic groups than eCO_2 samples (Fig. S3.2E). Such differences in richness at the alpha level resulted in significantly higher genetic richness at the gamma level (Fig. S3.2F). However, the evenness of genetic groups did not change significantly at both alpha and gamma level (Fig. S3.3EF). Those changes resulted in no significant changes of genetic diversity (measured by Shannon diversity) at the alpha level, but did at the gamma level ($P<0.05$), especially when > 2 million randomly selected sequences were analyzed (Fig. S3.4A-C).

Taxonomic diversity. Taxonomic diversity was analyzed based on OTUs of 16S rRNA amplicon sequencing and microbial species identified by shotgun metagenome sequencing approaches. The average richness of 16S OTUs in each aCO_2 sample was about 500 higher than that in eCO_2 samples (Fig. S3.2A). About 200 more sequenced microbial species were detected in aCO_2 samples than in eCO_2 samples by shotgun metagenome sequencing, with statistical significance p value of 0.08 (Fig. S3.2C). Significantly higher species richness in aCO_2 samples was also observed at the gamma level by 16S OTUs (Fig. S3.2B), but not by shotgun metagenome sequencing (Fig. S3.2D), which could be due to the limited microbial species identified by shotgun

metagenome sequencing. Significantly changed evenness of microbial species was neither observed at the alpha level, nor at the gamma level for shotgun metagenomes, but at the gamma level of 16S OTUs (Fig. S3.3A-D). As a result, higher alpha-diversity ($P<0.1$ for 16S OTUs) and significantly lower beta-diversity ($P<0.01$) of microbial species at aCO_2 were observed for both 16S OTUs and shotgun metagenomes (Fig. 3.2DE, Fig. S3.4DE), which led to significantly higher gamma-diversity in aCO_2 samples (Fig. 3.2F, Fig. S3.4F).

Functional diversity. Functional diversity was analyzed by functional groups identified by searching predicted proteins against the eggNOG database. Higher functional group richness in aCO_2 samples was observed at both alpha ($P<0.1$) and gamma level (Fig. S3.2GH). Similar to the taxonomic diversity, the evenness of functional groups did not significantly change at the alpha level, but at the gamma level (Fig. S3.3GH). Interestingly, the diversity of functional groups decreased at all three diversity components in response to eCO_2, with $P<0.1$ for alpha-, and $P<0.05$ for beta- or gamma-diversity (Fig. 3.2G-I).

Based on above results, it could be concluded that long-term eCO_2 decreased the overall microbial biodiversity as a result of functional, rather than taxonomic, convergence. This is evidenced by significantly decreased functional beta-diversity, and increased taxonomic and unweighted phylogenetic beta-diversity (Fig. 3.2BEH). The reduced functional beta-diversity indicated a more common set of functional gene groups in eCO_2 samples, while increased taxonomic/phylogenetic beta-diversity indicated a less overlapped set of microbial species in eCO_2 samples. Noticing that both richness and alpha diversity decreased in eCO_2 samples (Fig. S3.2, Fig. 3.2, Fig. S3.4),

the decreased functional beta- and increased taxonomic/phylogenetic beta-diversity suggested that eCO_2 selected microbial function rather than taxonomy. The increased taxonomic beta-diversity was not only evidenced by 16S rRNA sequencing, but also by microbial species derived from shotgun metagenomes (Fig. S3.4E). Furthermore, null model analysis by fixing alpha- and gamma-diversity (Chase, Kraft et al. 2011) suggested that both observed taxonomic/functional beta-diversity under aCO_2 and eCO_2 were significantly different from null random expectations (Table S3.1), confirming the robustness of our findings.

We then found that rare functional groups were mainly responsible for the decreased functional diversity in eCO_2 samples. In order to reveal the preference behavior of functional convergence under eCO_2, we first divided functional groups into two major categories (COGs–clusters of orthologous groups and NOGs—non-supervised orthologous groups) according to their origins in the eggNOG database (Muller, Szklarczyk et al. 2010). The NOGs were considered as rare functional groups in this study for their having 7.5 times more orthologous groups than COGs, but only consisting ~1/7 as many sequences as that in COGs (Fig. S3.5). Although significantly lower Shannon diversity was found in eCO_2 samples for both COGs and NOGs (Fig. S3.5EF), NOGs was mainly responsible for the decreased functional group richness (Fig. S3.5AB). This suggested that a functional convergence process under eCO_2 favored by dominant functional gene groups rather than rare ones, the latter of which may not be essential to maintain basic ecosystem functioning (Loreau, Naeem et al. 2001; Smith and Knapp 2003; Lyons, Brigham et al. 2005). At the functional category level, although it was not significant, consistently decreased Chao1 richness was

observed for all functional categories (Table S3.2). Among these, only functional group richness for "Secondary metabolites biosynthesis, transport and catabolism" and "Coenzyme transport and metabolism" were significantly lower in eCO_2 ($P \leq 0.02$), suggesting that the decreased functional diversity as an overall effect of all functional categories (Table S3.2).

Consequently, functional convergence resulted in altered diversity for taxonomic groups. Unlike the dominance of Verrucomicrobia in the tallgrass prairie ecosystem in Midwestern US (Fierer, Ladau et al. 2013), the soil microbial community in this grassland ecosystem was dominated by Actinobacteria and Proteobacteria, followed by relatively low abundances of Acidobacteria, Planctomycetes, Chloroflexi, Firmicutes, Verrucomicrobia, Bacteroidetes, and Cyanobacteria, as revealed by both 16S rRNA and shotgun metagenome sequencing (Fig. S3.6). Further investigation at the phylum level suggested that it was Proteobacteria and other rare phyla, not Actinobacteria, that were mainly responsible for the decreased taxonomic richness in eCO_2 (Fig. S3.7A-C). Taxonomic Shannon diversity decreased for both Actinobacteria and Proteobacteria, but increased for rare phyla (Fig. S3.7G-I), indicating a magnitudinal increase of evenness for these succeeded rare phyla (Fig. S3.7D-F).

The decreased microbial biodiversity was significantly correlated with soil NH_4^+. Long-term eCO_2 significantly stimulated both aboveground and root biomass production (Reich, Knops et al. 2001; Reich and Hobbie 2013), as well as ammonification rate in soil, but not soil moisture, pH or nitrification rate (Fig. S3.8). Positive correlations between aboveground biomass and ammonification rate could be observed ($P<0.1$) (Fig. 3.3A), suggesting a demand for more N in the form of NH_4^+ by

stimulated plant growth. Interestingly, significant or marginally significant negative correlations were observed between ammonification rate and functional richness ($P<0.1$), functional diversity ($P<0.05$), phylogenetic diversity ($P<0.05$), taxonomic richness ($P<0.05$) and taxonomic diversity ($P=0.14$) (Fig. 3.3 B-F), but not between plant biomass and diversity indices (Table S3.3). This suggested that the increased demand for NH_4^+ by stimulated plant growth was one of the major driving factors responsible for decreased microbial biodiversity.

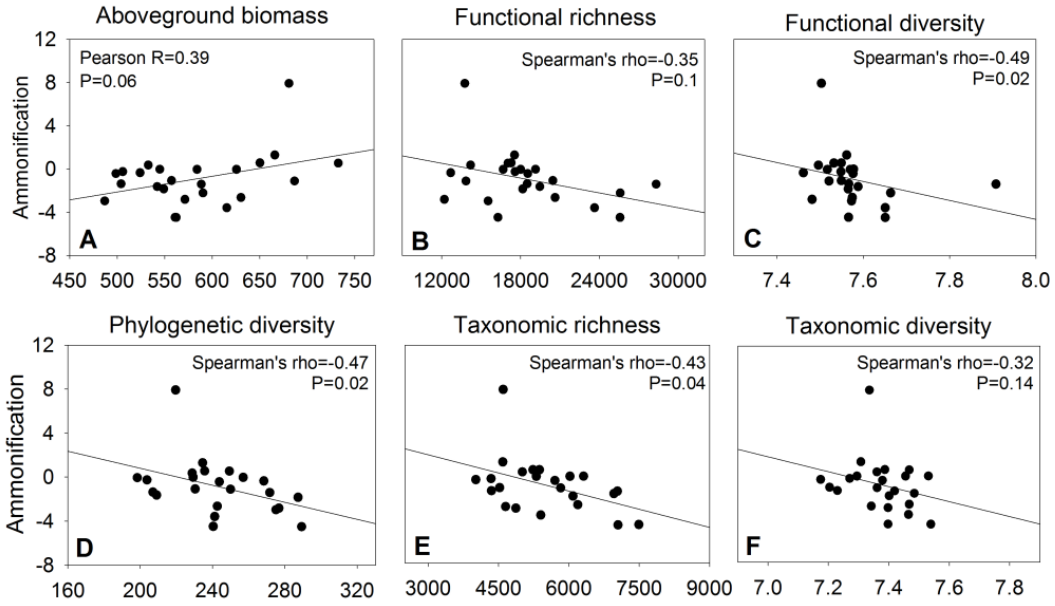


Fig. 3.3 Correlations between soil ammonification rate and aboveground plant biomass (A), functional richness (B), functional diversity (C), phylogenetic diversity (D), taxonomic richness (E), and taxonomic diversity (F). Soil ammonification rate was positively correlated with plant biomass, but negatively with diversity indices ($P<0.1$ except taxonomic diversity).

Notably, the correlation between soil NH_4^+ and microbial community was also evidenced by stimulated abundance of corresponding NH_4^+-producing gene families (Fig. S3.8). Specifically, relative abundance of gene families involved in organic decomposition (*ureC*), N_2 fixation (*nifH*), and dissimilatory NO_3^- reduction (*narHJ*,

napB, *nirB* and *nrfA*) that produce NH_4^+ from organic N, N_2, and NO_3^-/NO_2^- increased under eCO_2 ($P<0.1$). Interestingly, relative abundances for gene families that are responsible for assimilatory NO_3^- reduction, in which produced ammonium was used for microbial biomass synthesis, were not significantly changed (Fig. S3.10). The gene family *amoA* responsible for nitrification in soil remained unchanged (Fig. S3.10), consistent with our meta-analysis that nitrification rate was not significantly differed (Fig. S3.1A, Fig. S3.7C). More interestingly, the relative abundance of genes encoding glutamine synthetase that synthesizes glutamine from NH_4^+ and glutamate decreased as a result, indicating another potential mechanism of microbial communities to provide more NH_4^+ by reducing microbial NH_4^+ uptakes.

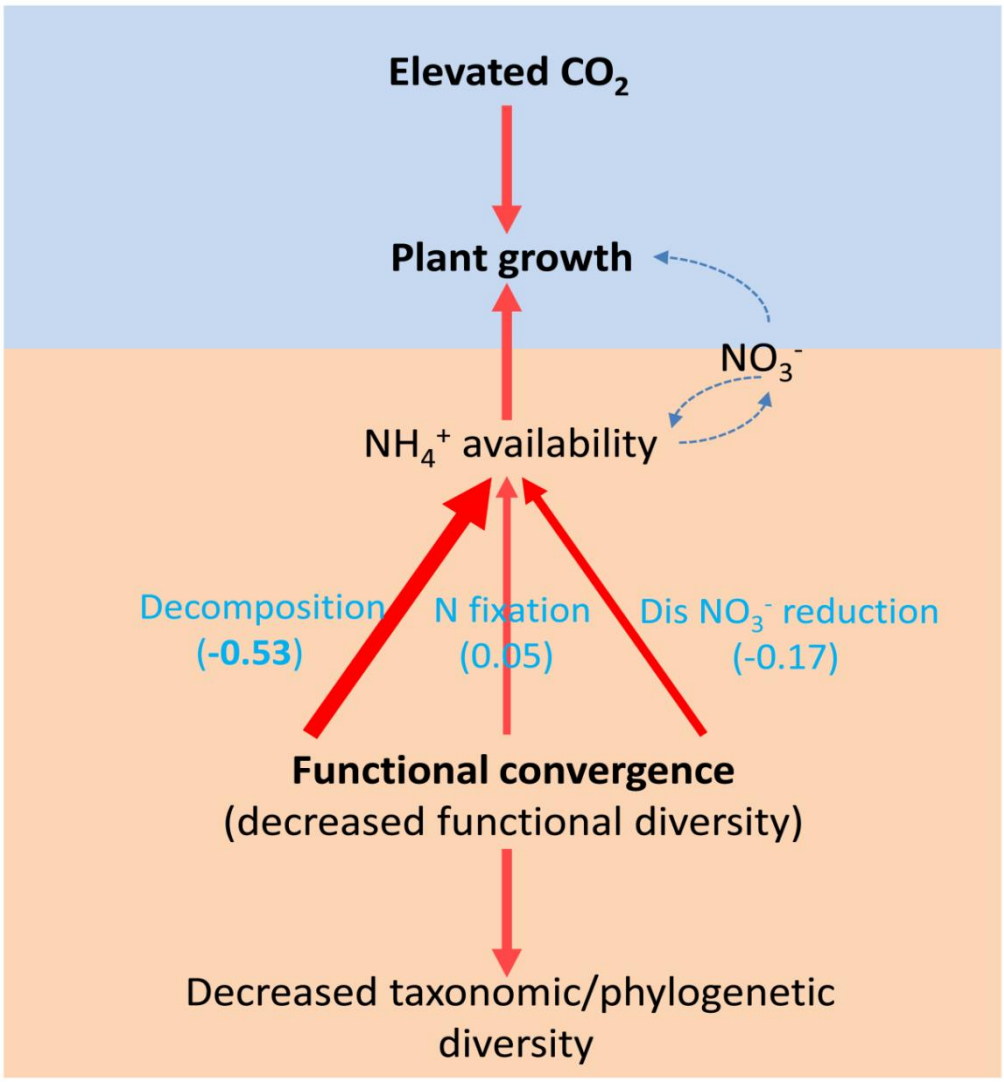


Fig. 3.4 A conceptual framework illustrating how long-term eCO_2 decreases microbial biodiversity. Long-term eCO_2 stimulated plant growth in grassland ecosystems, leading to progressive N limitation in soil. Microorganisms capable to produce NH_4^+ from various sources were favored, resulting in functional convergence of microbial communities. Decreased functional diversity as a result of functional convergence then led to decreased to taxonomic and phylogenetic diversity of microbial communities. Numbers in brackets indicate path coefficients as revealed by path analysis. Bolded numbers were significantly different from zero, based on bootstrap t-test.

In summary, we proposed a conceptual model to illustrate how long-term eCO_2 affects soil microbial biodiversity (Fig. 3.4). Long-term eCO_2 stimulated plant growth in grassland ecosystems (Reich, Knops et al. 2001; Norby, DeLucia et al. 2005; Luo,

Hui et al. 2006; Reich and Hobbie 2013), which proposed demands for more biologically available N in soil, a process termed as progressive N limitation (Hu, Chapin et al. 2001; Norby and Luo 2004; Reich, Hobbie et al. 2006; Norby, Warren et al. 2010; Reich and Hobbie 2013). As microorganisms are able to produce NH_4^+ from various sources, microbial species with such functions were favored (Zehr, Jenkins et al. 2003; Craine, Morrow et al. 2007), resulting in functional convergence of microbial communities. Of the three significantly increased NH_4^+-producing pathways, microbial decomposition is the major factor responsible for decreased functional diversity as revealed by path analysis. Decreased functional diversity as a result of functional convergence also led to decreased to taxonomic and phylogenetic diversity of microbial communities. As the microbial communities were favored by function, rather than taxonomy, decreased functional but increased taxonomic/phylogenetic beta diversity was observed, resulting in decreased microbial biodiversity in grassland ecosystems.

Our findings in this study revealed a fundamental but important mechanism that microbial communities respond to environmental perturbations by functional convergence, a process commonly found in macro ecosystems (Reich, Walters et al. 1997; Paruelo, Jobbágy et al. 1998; Meinzer 2003; Shaver, Street et al. 2007), but not yet well established for microbial communities. Our results also challenge many previous studies that showed no difference or increased microbial diversity under eCO_2 (Janus, Angeloni et al. 2005; Lipson, Wilson et al. 2005; Lipson, Blair et al. 2006; Lesaulnier, Papamichail et al. 2008; Castro, Classen et al. 2010; Dunbar, Eichorst et al. 2012), possibly owing to technical drawbacks and short-term experiments, for which rare species were not well captured and a balance between eCO_2, plant growth, nitrogen

limitation and microbial communities had not been established. Notably, although nitrogen limitation was found to be significantly correlated with decreased microbial biodiversity, other factors may also contribute importantly. This is solely because ecosystem responses to environmental perturbations are such a complex procedure that many other factors may not yet well explored (Norby and Luo 2004).

3.4 Materials and Methods

3.4.1 Site description and sample collection

The study was conducted within the BioCON (Biodiversity, CO_2 and N) experimental site located at the Cedar Creek Ecosystem Science Reserve in Minnesota, USA (45.4086° N, 93.2008° W). The long-term experiment was started in 1997 on a secondary successional grassland situated on a sandy outwash soil after removing the previous vegetation (Reich, Knops et al. 2001). The main BioCON field experiment has 296 (of a total of 371) evenly distributed plots (2 x 2 m) in six 20-meter diameter FACE (free air CO_2 enrichment) rings, three with aCO_2 concentrations, and three with CO_2 concentrations elevated by 180 μmol/mol (Lewin, Hendrey et al. 1994). In this study, 24 plots (12 from aCO_2, 12 from eCO_2, all with 16-species and no additional N supply) were used.

All of the 16 plant species used in this study are native or naturalized to the Cedar Creek Ecosystem Science Reserve, and can be classified into four functional groups: (i) four C_3 grasses (*Agropyron repens*, *Bromus inermis*, *Koeleria cristata*, *Poa pratensis*), (ii) four C_4 grasses (*Andropogon gerardii*, *Bouteloua gracilis*, *Schizachyrium scoparium*, *Sorghastrum nutans*), (iii) four N-fixing legumes (*Amorpha canescens*, *Lespedeza capitata*, *Lupinus perennis*, *Petalostemum villosum*), and (iv) four

non N-fixing herbaceous species (*Achillea millefolium*, *Anemone cylindrica*, *Asclepias tuberosa*, *Solidago rigida*). Plots were regularly manually weeded to remove unwanted species, although the 16 species plots used in this study require minimal weeding.

Bulk soil samples were taken in July, 2009 under ambient and eCO_2 conditions for microbial community analysis, and each sample was composited from five soil cores at a depth of 0-15 cm. All samples were immediately transported to the laboratory, frozen and stored at -80^{o}C for DNA extraction, PCR amplification, and 454 pyrosequencing.

3.4.2 Plant and soil property measurements

Plant biomass. The aboveground and belowground (0–20 cm) biomass were measured as previously described (Reich *et al.* 2001; Reich *et al.* 2006). A 10 x 100 cm strip was clipped at just above the soil surface, and all plant material was collected, sorted to live material and senesced litter, dried and weighed. Roots were sampled at 0–20 cm depth using three 5-cm diameter cores in the area used for the aboveground biomass clipping. Roots were washed, sorted into fine (< 1 mm diameter) and coarse classes and crowns, dried and weighed.

Soil physical properties. Soil pH and volumetric soil moisture were measured at depths of 0-17, 42-59, and 83-100 cm in a KCl slurry and with permanently placed TRIME Time Domain Reflectometry (TDR) probes (Mesa Systems Co., Medfield MA), respectively. The soil pH and moisture values measured at 0-17cm were used in this study.

Net N mineralization. Net N mineralization rates were measured concurrently in each plot for one-month in situ incubations with a semi-open core at 0-20 cm depth during midsummer of each year (Reich et al. 2001; Reich et al. 2006). Net N mineralization

rates were determined by the difference between the final and initial NH_4^+-N + NO_3^--N pool sizes determined with 1 M KCl extractions. Net ammonification was determined by the difference between the final and initial NH_4^+-N pool sizes. Net nitrification was determined by the difference between the final and initial NO_3^--N pool sizes.

3.4.3 DNA extraction, purification and quantification

Soil DNA was extracted by freeze-grinding mechanical lysis as described previously (Zhou, Bruns et al. 1996), and was purified using a low melting agarose gel followed by phenol extraction for all 24 soil samples collected. DNA quality was assessed by the ratios of 260/280 nm, and 260/230 nm using a NanoDrop ND-1000 Spectrophotometer (NanoDrop Technologies Inc., Wilmington, DE), and final soil DNA concentrations were quantified with PicoGreen (Ahn, Costa et al. 1996) using a FLUOstar Optima (BMG Labtech, Jena, Germany).

3.4.4 Shotgun metagenome sequencing and 16S rRNA gene amplicon sequencing

All 24 samples were subjected to shotgun metagenome sequencing by Roche 454 pyrosequencing approaches. Library construction and sequencing were carried out by Los Alamos National Lab (New Mexico, USA) using standard shotgun protocols.

A total of 23 samples instead of 24 were subjected to 16S rRNA gene amplification and MiSeq Illumina sequencing due to insufficient remaining DNA for one of the samples. PCR amplification was performed for the V4-V5 hypervariable regions of bacterial 16S rRNAs using the PCR primers, F515: GTGCCAGCMGCCGCGG, and R806: GGACTACHVGGGTWTCTAAT. A unique 12-mer barcode was added for each sample at the 5'-end of the forward primer. The barcode-primers were synthesized by Invitrogen (Carlsbad, CA) and used for the generation of PCR amplicons.

Quadruplicate 20 µl PCR reactions were performed as follows: 4 µl Promega GoTaq buffer, 0.5 µl GoTaq DNA polymerase, 1.5 µl Roche 25 mM MgCl2, 1 µl Invitrogen 10 mM dNTP mix, 1 µl of each primer (10 pmol μl^{-1}), 0.2 µl New England BioLabs 10 mg ml^{-1}BSA, 1 µl 10 ng 7 μl^{-1} template, and 9.8 µl H_2O. Cycling conditions were an initial denaturation of 94 ℃ for 3 min, 30 cycles of 94 ℃ for 1 min, 51 ℃ for 40 s, 72 ℃ for 1 min, and an final extension at 72 ℃ for 10 min. Replicates were pooled and gel purified using the Qiagen Gel Purification Kit following band excision. Products were further purified using the Qiagen PCR purification kit. After adapter ligation, amplicons were sequenced on a FLX 454 system (454 Life Sciences, Branford, CT) by Macrogen (Seoul, South Korea) using Lib-L kits and processed using the shotgun protocol.

3.4.5 Shotgun data preprocessing, gene prediction and annotation

A total of 18 454 run (12 full plate, 12 half plate) data was obtained. A total of 18,890,805 raw reads were obtained. Quality control for 454 shotgun sequences was carried out by the LUCY program (Chou and Holmes 2001) with minimum quality score of 21 and maximum error rate of 0.01, resulting in 17,096,024 high quality sequences. Gene prediction was carried out by FragGeneScan (Rho, Tang et al. 2010), which predicts high quality gene fragments from short, error-prone reads and overcomes homopolymer errors. A total of 17,578,392 genes were predicted by FragGeneScan. Gene groups for predicted genes were assigned by self-vs.-self BLAT (Kent 2002) approach with 90% identity cutoff over 30% alignment overlap. Function and taxonomy assignment of the predicted genes were carried out by searching protein sequences against eggNOG (Muller, Szklarczyk et al. 2010) and NCBI nr database, respectively. For function assignment, the best hit with eggNOG database was used.

And for taxonomy assignment, lowest common ancestors were assigned based on the best hits within 1/10 top e-value were used using the MEGAN program (Huson, Auch et al. 2007).

3.4.6 16S amplicon data processing and OTU identification

A total of 917,824 paired-end (2x150bp) MiSeq Illumina reads were obtained for 16S rRNA amplicons. Forward and reverse reads were joined by the FLASH program (Magoč and Salzberg 2011) with 10bp minimum overlap and allowing 2 mismatches. This resulted in 821,265 longer merged reads covering the primer region. Quality filtering, chimera removal and OTU clustering were carried out using the UPARSE pipeline (Edgar 2013), which is a recently developed approach that identifies highly accurate OTUs from amplicon sequencing data. Reads with expected errors >0.5 were discarded. The reads were then dereplicated, sorted, and clustered into candidate OTUs with an identity cutoff of 0.97. Chimeric OTUs were identified and removed by searching against the greengenes reference sequences (McDonald, Price et al. 2012). A total of 500,279 merged sequences were retained and clustered into 16,633 non-chimeric OTUs. Finally, qualified reads were mapped to OTU reference sequences for relative abundance calculation. Taxonomic assignment for OTUs was carried out by RDP classifier (Wang, Garrity et al. 2007). OTU representative sequences were aligned by the MUSCLE program (Edgar 2004), and a phylogenetic tree was built using FASTTREE (Price, Dehal et al. 2009).

3.4.7 Biodiversity definition and calculation

Biodiversity encompasses at least four dimensions: taxonomic, phylogenetic, genetic, and functional diversity. In order to comprehensively examine all these four dimensions

of biodiversity, different sequencing approaches were used. The following definitions and methods were used to analyze all dimensions of biodiversity.

Taxonomic diversity: diversity of microbial species. Taxonomic diversity was analyzed by 16S rRNA gene sequences, by defining operational taxonomic units (OTUs) based on sequence dissimilarity.

Phylogenetic diversity: diversity of evolutionary relationships of microbial species. Phylogenetic diversity can be analyzed by constructing phylogenetic trees from phylogenetic markers such as 16S rRNA genes.

Genetic diversity: the diversity of genetically inheritable regions. We defined genetic groups from predicted genes of shotgun metagenomes to analyze genetic diversity. Due to the short sequence length and not well assembly limitation of shotgun metagenomes, genetic groups were defined as predicted genes with 90% sequence identity over 30% overlaps.

Functional diversity: the diversity of functional traits of microbial community. Here we used functional groups to assess the functional diversity of belowground microbial communities. Functional groups were defined by searching predefined orthologous (COG/NOG) database, in which each orthologous group could be defined as a gene family.

For each dimension of biodiversity, all three components of diversity were analyzed, including alpha-, beta- and gamma-diversity. Richness and evenness were also analyzed at alpha- and gamma-level. Gamma-diversity was computed by combing all samples collected from the same treatment/control sites. In this study, we used Chao1 richness, Shannon evenness, and Shannon diversity to analyze richness,

evenness and diversity indices, respectively. Beta-diversity that represents the differences between two samples was analyzed by Bray-Curtis dissimilarity index. Phylogenetic diversity was analyzed by the mothur package, according to the method described in (Faith 1992). Phylogenetic beta diversity was assessed by both weigthed and unweighted UniFrac distance (Lozupone and Knight 2005).

Chapter 4: Fungal Communities Respond to Long-term Elevated CO_2 by Community Reassembly

4.1 Abstract

Fungal communities play key roles in Earth's ecosystems – they are important decomposers and parasites as well as symbionts of plants. Their community-level responses to eCO_2, one of the major global change factors impacting ecosystems, are not well understood. Using 28S rRNA gene amplicon sequencing and co-occurrence ecological network approaches, we analyzed the response of soil fungal communities in the BioCON experimental site in Minnesota, USA, in which a grassland ecosystem exposed to eCO_2 for 12 years. Long-term eCO_2 did not significantly change the overall fungal community structure and species richness, but significantly increased community evenness and diversity. Relative abundances of 119 OTUs (~ 27% of the total captured sequences) were changed significantly. Interestingly, significantly changed OTUs under eCO_2 were associated with decreased overall relative abundance of Ascomycota, but increased relative abundance of Basidiomycota. Co-occurrence ecological network analysis indicated that eCO_2 increased fungal community interactions, as evidenced by higher intermodular and intramodular connectivity and shorter geodesic distance. In contrast, decreased connections for dominant fungal species were observed in the eCO_2 network, and community reassembly of unrelated fungal species into highly connected dense modules was observed. Such changes in the network structure were significantly associated with altered soil and plant properties under eCO_2, especially with increased plant biomass and NH_4^+ availability. This study provides novel insights into how eCO_2 shapes soil fungal communities in grassland ecosystems.

Keywords: elevated CO_2; fungal community; microbial ecological network; community reassembly

4.2 Introduction

Fungi represent a significant portion of the microbial community in the Earth biosphere, with an estimated 1.5~5.1 million species in total (Hawksworth 2001; O'Brien, Parrent et al. 2005). They play various roles in terrestrial ecosystems, such as decomposers, parasites, and symbionts (Webster and Weber 2007). Understanding the fungal diversity, community structure and their responses to long-term eCO_2 in grassland ecosystems is an important issue in ecology and global change biology, but little is known about the impacts of eCO_2 on the diversity, composition, structure, and function of soil fungal communities due to the high diversity and uncultivable nature of most (>80%) soil fungi (Bridge and Spooner 2001).

Past studies have shown that eCO_2 significantly increases the plant productivity in grassland ecosystems, resulting in more C input to the soil (Reich, Knops et al. 2001; He, Xu et al. 2010; Langley and Megonigal 2010; Drake, Gallet-Budynek et al. 2011; Zak, Pregitzer et al. 2011; Reich and Hobbie 2013), while increased C input in turn significantly changed bacterial diversity, composition and structure, and increased the functional potential of bacterial communities for C degradation and nutrient cycling, though such effects differed across various ecosystems (Lesaulnier, Papamichail et al. 2008; Drigo, Van Veen et al. 2009; Blagodatskaya, Blagodatsky et al. 2010; Castro, Classen et al. 2010; Drigo, Pijl et al. 2010; Feng, Simpson et al. 2010; He, Xu et al. 2010; Deng, He et al. 2012; Hayden, Mele et al. 2012; He, Piceno et al. 2012; Drigo, Kowalchuk et al. 2013). By contrast, fungal biomass and relative abundance did not change significantly under eCO_2 in these studies (Chung, Zak et al. 2007; He, Xu et al. 2010). Previous studies of fungal responses to eCO_2 were mainly carried out using

approaches such as phospholipid fatty-acid analysis (PLFA), denaturing gradient gel electrophoresis (DGGE), extracellular enzyme assays, and clone library analysis (Chung, Zak et al. 2007; Parrent and Vilgalys 2007; Drigo, Kowalchuk et al. 2008; Drigo, Van Veen et al. 2009; Castro, Classen et al. 2010; He, Xu et al. 2010), and mostly focused on mycorrhizal fungi (Alberton, Kuyper et al. 2005; Drigo, Pijl et al. 2010; Antoninka, Reich et al. 2011; Cheng, Booker et al. 2012), which have major influences on plant biodiversity and productivity (van der Heijden, Klironomos et al. 1998). Those previous studies were focused on fungal C degradation, N cycling, and interactions with plants (Cheng, Booker et al. 2012; Phillips, Meier et al. 2012; Verbruggen, Veresoglou et al. 2013); however, knowledge about fungal community-level responses to eCO_2 is still limited, though some efforts have been made recently (Castro, Classen et al. 2010; Edwards and Zak 2011; Weber, Vilgalys et al. 2013).

Microorganisms, including bacteria, archaea, viruses, fungi and protists, interact with each other in soil to form complex interactive networks (Faust and Raes 2012). Using ecological network approaches, co-occurrence ecological networks of microbial communities can be constructed and analyzed (Zhou, Deng et al. 2010; Steele, Countway et al. 2011; Zhou, Deng et al. 2011; Barberan, Bates et al. 2012; Faust, Sathirapongsasuti et al. 2012). For example, a global ecological network analysis of the human microbiome revealed 3,005 co-occurrence and co-exclusion relationships among 197 clades occurring throughout the human microbiome (Faust, Sathirapongsasuti et al. 2012). For environmental perturbation impacts on microbial network structures, previous studies showed that eCO_2 significantly impacted soil bacterial/archaeal community networks in a grassland ecosystem, and that significantly different network

structures and increased network complexity were observed in response to eCO_2 (Zhou, Deng et al. 2010; Zhou, Deng et al. 2011). However, the network interactions of large soil microorganisms (e.g., fungi, protists) and their responses to eCO_2 have not yet been characterized. Therefore, much can be learned by exploring interactions within fungal communities under eCO_2 using molecular ecological network analysis (MENA) (Deng, Jiang et al. 2012).

In this study, we aimed to comprehensively survey the fungal community diversity and examine their changes in composition, structure, and co-occurrence interactions in response to eCO_2 in a grassland soil ecosystem. The following hypotheses will be tested: (1) Stimulated plant biomass and changed soil properties as a result of eCO_2 would significantly affect the fungal community structure and diversity; (2) Such stimulated plant growth and higher demand for biologically available N would promote fungal communities to form more effective collaborations for organic decomposition, resulting in reassembled communities. To test the above hypotheses, we examined the response of fungal communities to long-term eCO_2 in the BioCON experimental site, a 12-year CO_2 manipulation in temperate grassland in central Minnesota, USA, by sequencing 28S rRNA gene amplicons, and comparing fungal community co-occurrence networks under ambient aCO_2 and eCO_2. Our results indicated that eCO_2 increased the fungal species richness and fungal community interactions, but decreased the community diversity and connections for dominant species. Such changes were significantly associated with soil and plant properties. This study provides novel insights into how eCO_2 shapes soil fungal communities in

grassland ecosystems, improving our understanding of the effects of eCO_2 on soil fungal communities.

4.3 Materials and Methods

4.3.1 Site description and sample collection

The same samples collected from the BioCON experimental site were used. Please refer to Chapter 3 for more detailes.

4.3.2 DNA extraction, purification and quantification

Soil DNA was extracted by freeze-grinding mechanical lysis as described previously (Zhou, Bruns et al. 1996), and was purified using a low melting agarose gel followed by phenol extraction for all 24 soil samples collected. DNA quality was assessed by the ratios of 260/280 nm, and 260/230 nm using a NanoDrop ND-1000 Spectrophotometer (NanoDrop Technologies Inc., Wilmington, DE), and final soil DNA concentrations were quantified with PicoGreen (Ahn, Costa et al. 1996) using a FLUOstar Optima (BMG Labtech, Jena, Germany).

4.3.3 PCR amplification and 454 pyrosequencing

A total of 23 samples instead of 24 were subjected to 454 pyrosequencing due to insufficient remaining DNA for one of the samples. Amplification was performed using a fungal 28S rRNA gene primer pair with the forward primer LR3: ACCCGCTGAACTTAAGC, and the reverse primer LR0R: CCGTGTTTCAAGACGGG, whose products are expected to be approximately 625-bp (Liu, Porras-Alfaro et al. 2012). A unique 8-mer barcode was added for each sample at the 5'-end of the forward primer. The barcode-primers were synthesized by Invitrogen (Carlsbad, CA) and used for the generation of PCR amplicons. Quadruplicate 20 µl

PCR reactions were performed as follows: 4 µl Promega GoTaq buffer, 0.5 µl GoTaq DNA polymerase, 1.5 µl Roche 25 mM MgCl2, 1 µl Invitrogen 10 mM dNTP mix, 1 µl of each primer (10 pmol μl^{-1}), 0.2 µl New England BioLabs 10 mg ml^{-1}BSA, 1 µl 10 ng 7 μl^{-1} template, and 9.8 µl H_2O. Cycling conditions were an initial denaturation of 94 ℃ for 3 min, 30 cycles of 94 ℃ for 1 min, 51 ℃ for 40 s, 72 ℃ for 1 min, and an final extension at 72 ℃ for 10 min. Replicates were pooled and gel purified using the Qiagen Gel Purification Kit following band excision. Products were further purified using the Qiagen PCR purification kit. After adapter ligation, amplicons were sequenced on a FLX 454 system (454 Life Sciences, Branford, CT) by Macrogen (Seoul, South Korea) using Lib-L kits and processed using the shotgun protocol.

4.3.4 Data analysis

Raw pyrosequencing reads were extracted from the *sff* file using the *sffinfo* tool from Roche 454. Two files, a *fasta* file containing the sequence and a *qual* file containing the quality information, were generated and then converted into a *fastq* file using the python script "faqual2fastq2.py" that comes with the UPARSE pipeline (Edgar 2013). The quality filtering, chimera removal and OTU clustering were carried out using the UPARSE pipeline (Edgar 2013), which is a recently developed approach that identifies highly accurate OTUs from amplicon sequencing data. Only the reads with perfectly matched barcodes and maximum of 2 primer mismatches were kept for further analysis. Barcodes and primers were deleted from reads. The remaining reads were then truncated to 250 bp, and reads with expected errors >0.5 were discarded. The reads were then dereplicated, sorted, and clustered into candidate OTUs with an identity cutoff of 0.97. Chimeric OTUs were identified and removed by searching against the

LSU reference sequences downloaded from the Silva database (release_111) (Pruesse, Quast et al. 2007). Finally, qualified reads were mapped to OTU reference sequences for relative abundance calculation.

Taxonomic assignment for OTUs was carried out by RDP classifier using the fungal LSU training dataset (Liu, Porras-Alfaro et al. 2012). Reference OTUs were aligned by the MUSCLE program (Edgar 2004), and a phylogenetic tree was built using FASTTREE (Price, Dehal et al. 2009). The tree was then rooted to an outlier OTU with taxonomy information assigned as "leaf". Significance tests for different taxonomic groups and OTUs were performed by response ratio analysis (Hedges, Gurevitch et al. 1999) at 95% confidence interval. Nonmetric multidimensional scaling (NMDS) analysis was performed in R using the package vegan. Species richness, evenness, and diversity indices were calculated by the Mothur package (Schloss, Westcott et al. 2009), with rarefaction analysis of 1000 bootstrap random sampling iterations and 0.1% incremental sampling efforts.

4.3.5 Co-occurrence ecological network construction and analysis

Fungal co-occurrence ecological networks were constructed and analyzed using the online MENA pipeline, which implements random matrix theory for threshold identification (Deng, Jiang et al. 2012). In order to construct highly confident fungal co-occurrence ecological networks for comparative analysis, several different approaches were applied. First, we used an RMT-based approach to identify a proper threshold for pairwise Pearson correlation coefficient values between OTUs. The RMT identifies the threshold by observing a transition point of nearest-neighbor spacing distribution of eigenvalues from Gaussian to Poisson distribution, which are two universal extreme

distributions (Zhou, Deng et al. 2010). The RMT-based approach is a reliable and robust tool for network construction and has been successfully applied to construct various networks, including gene regulatory networks (Luo, Yang et al. 2007; Yang, Harris et al. 2008; Zhou, He et al. 2010; Lin, Song et al. 2011; Lin, Ji et al. 2013), functional molecular ecological networks (Zhou, Deng et al. 2010), and phylogenetic molecular ecological networks (Zhou, Deng et al. 2011). Second, the same cutoff of 0.78 was applied to construct co-occurrence networks for fungal communities at aCO_2 and eCO_2, with the purpose of comparing between different networks. Since a smaller threshold will result in less reliable and larger networks with more nodes, the same cutoff could effectively eliminate imbalances in network comparisons. Third, only OTUs presented in at least 6 samples were used for Pearson correlation coefficient calculations and zero was filled in for missing values for OTUs in paired samples. This made the correlation coefficient between two OTUs more statistically reliable. Finally, in order to statistically compare the constructed networks, 100 randomly generated networks were created for both aCO_2 and eCO_2 with the same OTUs in each corresponding network. Network topological properties such as small world, scale-free, and modularity (Wang and Chen 2003) were then compared between the constructed networks and these random networks. Ecological networks were visualized by Cytoscape (Smoot, Ono et al. 2011).

4.3.6 Linking community structure and network topology with soil and plant properties

To analyze if the changed fungal community structure and network topology were correlated with soil and plant properties, Mantel tests that calculate the correlation between two matrices were performed. A total of eight soil and plant properties,

including soil moisture (0~17cm); pH; mid-season in situ net nitrification, ammonification, and N mineralization rate; total plant biomass; and mid-season extractable soil NH_4^+ and NO_3^- concentrations were collected and analyzed. Euclidean distance was used to construct dissimilarity matrices for both OTU-based tables (community structure, network topology) and environmental variable(s). For Mantel tests of correlations between network topology and soil and plant properties, the correlation between OTU significance (calculated by OTU relative abundance and soil and plant properties) and node connectivity was examined. More details could be found in (Deng, Jiang et al. 2012).

4.4 Results

4.4.1 CO_2 effects on soil and plant characteristics

Because soil and plant properties are directly related with belowground microbial community, the CO_2 effects on soil moisture, pH, mid-season *in situ* net nitrification, ammonification, and N mineralization rate, and plant biomass were analyzed. No significant change of mid-season net soil nitrification rate was found between aCO_2 and eCO_2 samples. However, the net soil ammonification rate was significantly ($P<0.05$) higher in eCO_2 samples than that in aCO_2 samples, resulting in moderately significantly ($P<0.1$) higher net N mineralization rate (Fig. 4.1A). The total plant biomass, as expected, also increased significantly ($P<0.05$) as a result of eCO_2 and higher soil N availability (Fig. 4.1B). The proportional soil moisture and pH, however, did not change significantly (Fig. S4.1), suggesting that the increased plant biomass and soil ammonification could be the major factors affecting belowground microbial communities, including fungal communities.

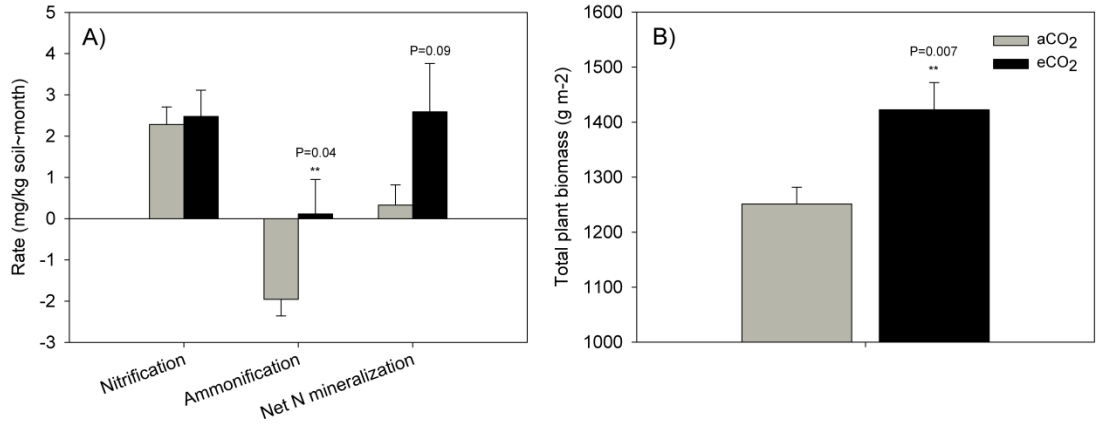


Fig. 4.1 eCO_2 effects on soil nitrogen (A) and total plant biomass (B). Both soil ammonification rate and plant growth were significantly stimulated after 12 years CO_2 treatment.

4.4.2 Sequence summary

Using 454 pyrosequencing, a total of 402,265 raw sequences of 28S rRNA gene amplicons were obtained with an average length of 477 bp for all 23 samples. A total of 339,048 reads (154,541 for aCO_2 samples, and 184,507 for eCO_2 samples) were then clustered into 1,975 OTUs after quality trimming, dereplication, clustering, and chimera removal by the UPARSE pipeline, with an OTU identity cutoff of 97%. Of the identified 1,975 OTUs, 407 were found to be singletons. Taxonomic assignment by RDP classifier showed 1,744 OTUs covering 97.9% qualified reads were fungal 28S rRNAs, and the remaining 231 were assigned to Eukaryota incertae sedis, but with 〈50% bootstrap confidence. Of these, 734 OTUs belonged to Ascomycota, 326 to Chytridiomycota, 298 to Basidiomycota, 96 to Blastocladiomycota, 53 to Glomeromycota, 41 to Neocallimastigomycota, and 2 to Zygomycota.

4.4.3 Long-term eCO_2 did not change the overall fungal community structure, but increased their diversity

The overall community structure between aCO_2 and eCO_2 samples was not significantly different, as revealed by three different non-parametric multivariate analysis methods (adonis: R=0.04, P=0.5; ANOSIM: R=0.03, P=0.71; MRPP: δ=0.55, P=0.52). This is also reflected by NMDS ordination analysis that no clear separation of aCO_2 samples from eCO_2 samples was found (Fig. S4.2).

To understand how long-term eCO_2 affects the fungal community diversity, the species richness and community diversity were analyzed by Chao1 index, Shannon evenness, Shannon diversity, and phylogenetic diversity. Shannon diversity treats each OTU as an independent entity (Hill 1973), and phylogenetic diversity (Vane-Wright, Humphries et al. 1991) considers the phylogenetic relationship among different OTUs. Owing to the close relationship between diversity indices and sequencing depth, a random subsampling effort of 6029 reads per sample was carried out by excluding four samples (two aCO_2 and two eCO_2) with less than 3000 reads. As a result, long-term eCO_2 did not significantly change the overall fungal species richness either, because 95% confidence intervals were clearly overlapped (Fig. S4.3A). However, the overall phylogenetic diversity (Fig. S4.3B) and taxonomic diversity (measured by Shannon diversity) (Fig. S4.3D) increased significantly, suggesting increased evenness of phylogenetically distant fungal species (Fig. S4.3C).

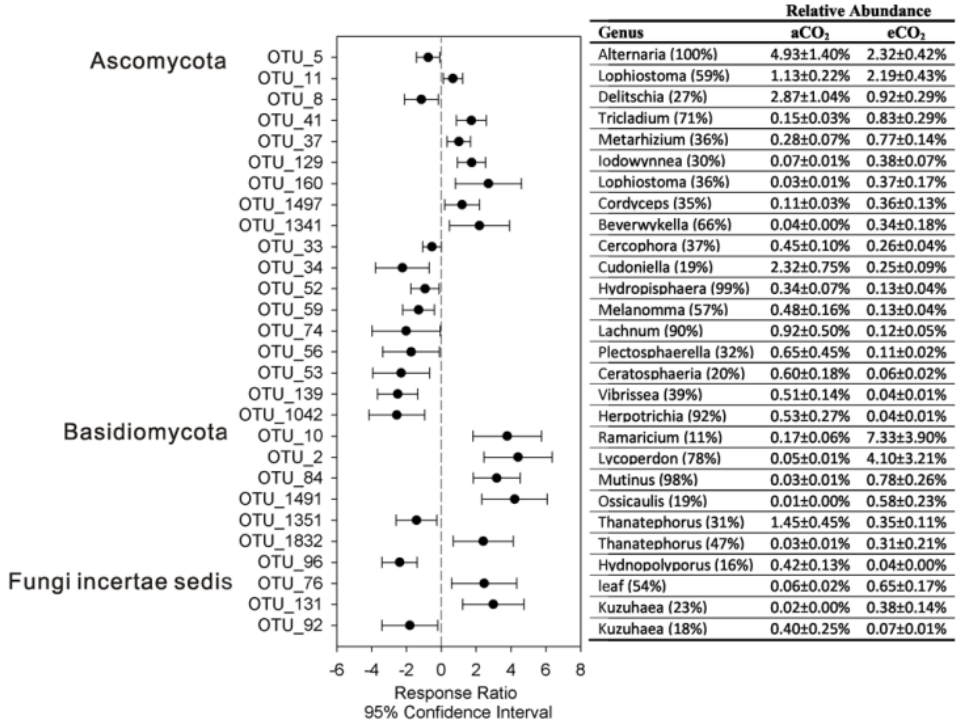
4.4.4 The composition of fungal community in grassland soil ecosystems

With a 50% bootstrap confidence cutoff, the fungal community in this grassland soil was dominated by Ascomycota (81% and 77% of sequences for aCO_2 and eCO_2,

respectively) and Basidiomycota, (11% and 14% of sequences for aCO_2 and eCO_2, respectively), followed by 1% Fungi incertae sedis, 0.25% Chytridiomycota, 0.05% Blastocladiomycota and 0.03% Glomeromycota at the phylum level. About 7% and 8% sequences in aCO_2 and eCO_2 samples could not be assigned to any phylum at 50% bootstrap confidence (Fig. S4.4A). At the order level, the most dominant fungal orders were Pleosporales (27.5%), Capnodiales (10.2%), Sordariales (7.5%), Hypocreales (5.4%), Helotiales (4.6%), Agaricales (5.2%), Thelebolales (3.2%), Chaetothyriales (2.6%), Cantharellales (2.3%), Coniochaetales (1.4%), Magnaporthales (1.4%), Xylariales (1.3%), Pezizales (1.3%), and Thelephorales (0.8%) (Fig. S4.4B). These 14 dominant fungal orders accounted for 74.7% of the total 28S rRNA sequences obtained. No significant differences were found for the relative abundances of the above dominant fungal phyla and orders between aCO_2 and eCO_2 samples.

Of the total 1,975 OTUs, the top 20 most abundant OTUs accounted for 50.3% and 50.2% of the total sequences for aCO_2 and eCO_2 samples, respectively. Three OTUs (OTU_1, OTU_3, and OTU_6) had ≥5% relative abundance in both aCO_2 and eCO_2 samples, and were assigned to genera *Davidiella* (70% bootstrap confidence), *Corynespora* (77% bootstrap confidence), and *Didymella* (47% bootstrap confidence), respectively. Relative abundance of a total of 119 OTUs significantly changed between aCO_2 and eCO_2 samples. Among these, 28 had ≥ 0.3% average relative abundance in aCO_2 or eCO_2 samples, including 18 from Ascomycota, 7 from Basidiomycota, and 3 from Fungi incertae sedis (Fig. 4.2). A total of 14 out of these 28 OTUs were found with significantly increased relative abundance in eCO_2 samples, including 7 Ascomycota OTUs, 5 Basidiomycota OTUs, and 2 incertae sedis fungal OTUs. Of the

14 OTUs with significantly decreased relative abundance in eCO_2, 11 were from Ascomycota, 2 from Basidiomycota, and one from incertae sedis fungus (Fig. 4.2). Interestingly, these significantly changed OTUs under eCO_2 were associated with decreased overall relative abundance of Ascomycota (11.5% in eCO_2 vs. 19.3% in aCO_2), but increased relative abundance of Basidiomycota (14.3% in eCO_2 and 3.5% in aCO_2). The top three most abundant OTUs with significantly increased relative abundance at eCO_2 were OTU_10 (*Ramaricium*, 11% bootstrap confidence), OTU_2 (*Lycoperdon*, 78% bootstrap confidence), and OTU_11 (*Lophiostoma*, 59% bootstrap confidence). The top four most abundant OTUs with significantly decreased relative abundance were OTU_5 (*Alternaria*, 100% bootstrap confidence), OTU_8 (*Delitschia*, 27% bootstrap confidence), OTU_34 (*Cudoniella*, 19% bootstrap confidence), and OTU_1351 (*Thanatephorus*, 31% bootstrap confidence). These 28 significantly changed OTUs accounted for 24.2% and 19.01% of the total captured sequences in eCO_2 and eCO_2 samples, respectively, while the total 119 significantly changed OTUs accounted for 27.9% and 24.8% of aCO_2 and eCO_2 samples, respectively (Fig. 4.2).



	Relative Abundance	
Genus	aCO_2	eCO_2
Alternaria (100%)	4.93±1.40%	2.32±0.42%
Lophiostoma (59%)	1.13±0.22%	2.19±0.43%
Delitschia (27%)	2.87±1.04%	0.92±0.29%
Tricladium (71%)	0.15±0.03%	0.83±0.29%
Metarhizium (36%)	0.28±0.07%	0.77±0.14%
Iodowynnea (30%)	0.07±0.01%	0.38±0.07%
Lophiostoma (36%)	0.03±0.01%	0.37±0.17%
Cordyceps (35%)	0.11±0.03%	0.36±0.13%
Beverwykella (66%)	0.04±0.00%	0.34±0.18%
Cercophora (37%)	0.45±0.10%	0.26±0.04%
Cudoniella (19%)	2.32±0.75%	0.25±0.09%
Hydropisphaera (99%)	0.34±0.07%	0.13±0.04%
Melanomma (57%)	0.48±0.16%	0.13±0.04%
Lachnum (90%)	0.92±0.50%	0.12±0.05%
Plectosphaerella (32%)	0.65±0.45%	0.11±0.02%
Ceratosphaeria (20%)	0.60±0.18%	0.06±0.02%
Vibrissea (39%)	0.51±0.14%	0.04±0.01%
Herpotrichia (92%)	0.53±0.27%	0.04±0.01%
Ramaricium (11%)	0.17±0.06%	7.33±3.90%
Lycoperdon (78%)	0.05±0.01%	4.10±3.21%
Mutinus (98%)	0.03±0.01%	0.78±0.26%
Ossicaulis (19%)	0.01±0.00%	0.58±0.23%
Thanatephorus (31%)	1.45±0.45%	0.35±0.11%
Thanatephorus (47%)	0.03±0.01%	0.31±0.21%
Hydnopolyporus (16%)	0.42±0.13%	0.04±0.00%
leaf (54%)	0.06±0.02%	0.65±0.17%
Kuzuhaea (23%)	0.02±0.00%	0.38±0.14%
Kuzuhaea (18%)	0.40±0.25%	0.07±0.01%

Fig. 4.2 Response ratio (eCO_2 vs. aCO_2) analysis of fungal OTU changes in response to eCO_2. Only the top 28 most abundant OTUs with relative abundance ≥0.3% in aCO_2 or eCO_2 were plotted. Error bar symbols plotted at the right of dashed line indicated increased relative abundances at eCO_2, while error bar symbols plotted at the left of dashed line indicated decreased relative abundances at eCO_2. The genus information as well as actual relative abundance with standard error was also listed.

4.4.5 The co-occurrence networks of fungal communities and their responses to eCO_2

Similar to other microorganisms (Faust and Raes 2012), fungi do not exist alone and interact with each other to form complex ecological interaction networks. In order to understand how fungal communities assemble and whether long-term eCO_2 affects the fungal community interaction, co-occurrence ecological networks were constructed for aCO_2 and eCO_2 fungal communities. Using the random matrix theory approach, a Pearson correlation coefficient cutoff of 0.78 was determined for network construction. Network comparisons were then carried out at both global level and sub-network level of selected nodes. The constructed aCO_2 fungal network contained 271 nodes (OTUs),

647 links and 19 modules (12 with ≥ 3 nodes), with an average connectivity of 4.78, average geodesic distance of 6.0, and modularity of 0.86, while the eCO_2 network had 226 nodes, 600 links and 13 modules (9 with ≥3 nodes), with an average connectivity of 5.31, average geodesic distance of 5.34, and modularity of 0.80 (Table S4.1, Fig. S4.5). Although the eCO_2 network contained fewer nodes and links, it is more complex than the aCO_2 network regarding the average connectivity, geodesic distance, and modularity, and the Student t-test showed that the average geodesic distance and modularity were significantly smaller in the eCO_2 network, suggesting that the nodes in eCO_2 network were more intensely connected with each other (Table S4.1). Both networks were dominated by OTUs from Ascomycota, which is also the dominant phylum in the fungal community (Fig. S4.5). For the 9 modules with ≥3 nodes in the eCO_2 network, 92 intermodular connections that linked different modules together were observed. In contrast, only 41 intermodular links were found for the 12 modules in the aCO_2 network. Since modules are composed of different fungal OTUs/species that have higher connectivity with within module members than outside module members, these modules could be regarded as putative microbial ecological niches (Zhou, Deng et al. 2010). Thus increased intermodular connections might indicate increased interactions between different fungal community "niches". In addition, more negative links were found in the eCO_2 network than in the aCO_2 network (47 in eCO_2 vs. 37 in aCO_2), suggesting that eCO_2 may also have increased competition among fungal species.

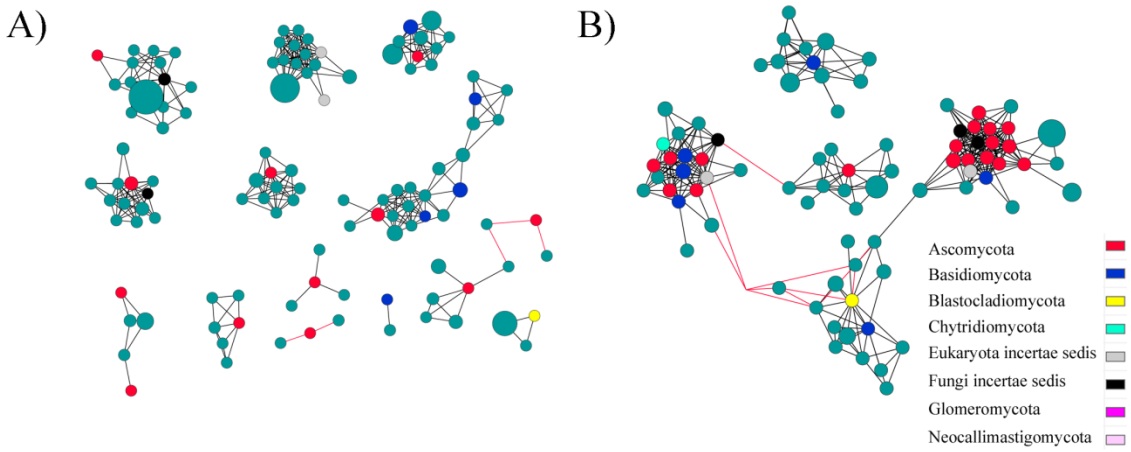


Fig. 4.3 Community reassembly of sparsely distributed OTUs in the aCO_2 network (A) into highly connected dense modules in the eCO_2 network (B). Colored nodes were the OTUs involved in community reassembly. Teal nodes were the first neighbor of yellow nodes. Different colors refer to different fungal phyla.

In addition to our comparisons of global network topological parameters between aCO_2 and eCO_2 networks, we also analyzed the effect of eCO_2 on sub-networks of fungal communities. Interestingly, 31 nodes that were sparsely distributed in 13 independent modules in the aCO_2 network (Fig. 4.3A) formed five dense modules with high connectivity in the eCO_2 network (Fig. 4.3B). Such interesting community reassembly process was not as obviously observed in the converse manner, i.e. dense aCO_2 modules did not separate into sparse individual nodes in eCO_2 networks. Of the 31 nodes, 27 were connected to each other in two major modules, and four of the five sub-modules were connected to another one (Fig. 4.3B). This was also consistent with the global observation that eCO_2 increased the intermodular connections. However, increased connectivity was not found for all the nodes in the eCO_2 network. For example, in the aCO_2 network, seven OTUs with high relative abundances ($\geq 2\%$) were connected with 37 first neighbors and formed relatively complex sub-networks with 145 links (Fig. 4.4A). In the eCO_2 network, although 5 of the 7 OTUs remained as the most

abundant OTUs in the network, they only connected to 20 neighbors with 31 links (Fig. 4.4B), resulting in much simpler network structure. The results indicated that long-term eCO_2 decreased the connectivity of OTUs with high relative abundances, but increased the connectivity for OTUs with lower relative abundances.

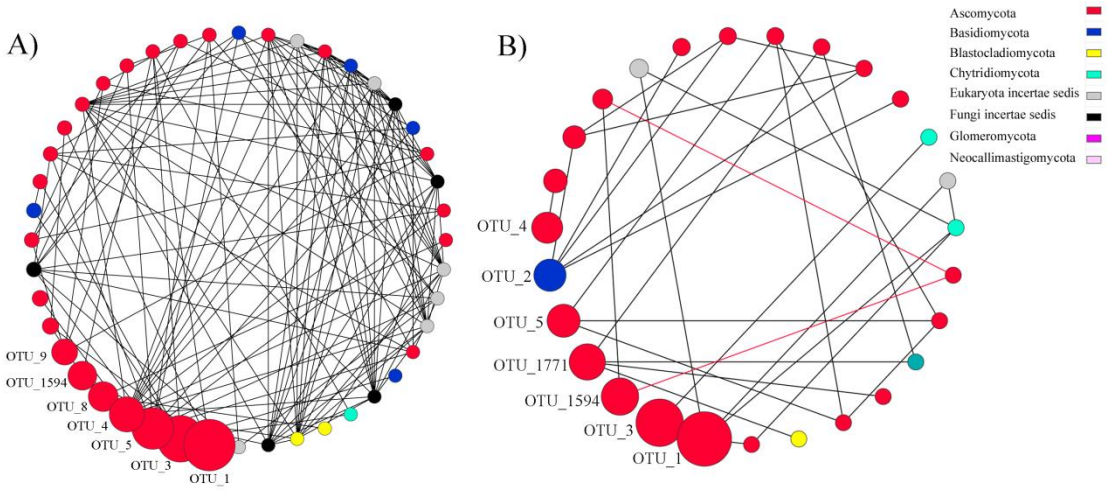


Fig. 4.4 The sub-network of top 7 most abundant OTUs and their first neighbor nodes in the aCO_2 (A) and eCO_2 (B) networks. Each node represents an OTU, which would be regarded as a fungal species. The size of nodes represents relative abundance of OTUs. Each link connects two OTUs. Grey links means positive connections, and red means negative connections. Different colors refer to different fungal phyla. The OTUs with top relative abundances were marked with OTU ids.

4.4.6 Linking fungal community structure and network topology with soil and plant properties

To determine if the fungal community structure and changed co-occurrence network topology were associated with soil and plant properties, Mantel test was performed. The relationships between community structure and soil moisture (0~17 cm), pH, mid-season in situ net nitrification, ammonification, and N mineralization rate, and total plant biomass were analyzed (Table 4.1). Consistent with our dissimilarity testing that the community structure did not differ from each other, no significant (R=-0.259, P =

0.981) correlation was observed between the overall community structure and the overall soil and plant properties, nor with any single soil and plant properties (Table 4.1).

Table 4.1 Mantel analysis of the relationships between the overall fungal community structure, co-occurrence network topology and individual soil properties.

Soil Properties	Community Structure		Network Topology			
			aCO2		eCO2	
	r_M	P	r_M	P	r_M	P
Soil moisture (0~17cm)	0.044	0.325	-0.002	0.516	-0.08	1
pH	-0.091	0.728	-0.052	0.970	-0.07	0.982
Mid-season *in situ* net nitrification rate	-0.178	0.898	0.0003	0.457	-0.07	0.984
Mid-season *in situ* net ammonification rate	-0.206	0.973	0.037	0.167	0.359	**0.001**
Mid-season *in situ* net N Mineralization rate	-0.256	1	0.100	**0.01**	0.077	**0.03**
Total plant biomass	0.022	0.392	-0.021	0.272	0.063	**0.07**

For Mantel tests between network topology and soil and plant properties, the trait-based OTU significance measure was used to determine a common group of soil and plant properties important to the network structure (Zhou, Deng et al. 2011). Mantel test of network topology and each soil and plant property suggested that soil ammonification and total plant biomass to be the major factors responsible for changed network topology (Table 4.1). There was a significant correlation between node connectivity and OTU significance of the selected soil variables based on all nodes (OTUs) with P = 0.001 (Table 4.2). Not all nodes in the network showed significant correlations with soil ammonification rate and plant biomass. Significant correlations mainly occurred for OTUs belonging to Ascomycota (P = 0.001), Basidiomycota (P=0.04) and incertae sedis fungi (P = 0.006). All the four major orders in Ascomycota, including Sordariomycetes (P = 0.001), Dothideomycetes (P = 0.001), Leotiomycetes (P = 0.02), and Lecanoromycetes (P=0.02) were significantly correlated with soil and plant properties (Table 4.2). For the aCO_2 co-occurrence network, as expected, no significant correlations were found between the node connectivity and OTU significance of the

selected soil and plant variables except Sordariomycetes, Fungi incertae sedis and Blastocladiomycota (Table 4.2)**.** The above results suggest that the changes of the co-occurrence fungal ecological network topology were significantly associated with increased soil ammonification rate and plant biomass under long-term eCO_2, and that OTUs belonging to Ascomycota were mainly responsible for such changes.

Table 4.2 Mantel test on network connectivity vs. the OTU significances of soil geochemical variables[a]

Phylogeny	aCO_2			eCO_2		
	#nodes	r_M[b]	P[c]	#nodes	r_M	P
All OTUs	271	0.037	0.146	226	0.363	**0.001**
Ascomycota	135	0.052	0.133	130	0.43	**0.001**
Sordariomycetes	46	0.209	**0.026**	47	0.463	**0.001**
Dothideomycetes	40	0.07	0.167	42	0.471	**0.001**
Leotiomycetes	16	-0.012	0.394	14	0.323	**0.02**
Lecanoromycetes	12	0.167	0.24	14	0.499	**0.02**
Basidiomycota	58	-0.011	0.493	34	0.217	**0.038**
Eukaryota incertae sedis	25	-0.071	0.671	21	0.268	0.06
Fungi incertae sedis	21	0.180	**0.035**	16	0.483	**0.006**
Chytridiomycota	18	0.205	0.086	11	0.036	0.424
Blastocladiomycota	8	0.575	**0.003**	7	-0.086	0.423
Glomeromycota	4	-0.451	0.963	5	0.542	0.052

[a]mid-season in situ net ammonification and total plant biomass were selected for their significant contribution to network topology differences
[b]Correlation based on Mantel test.
[c]The significance (probability) of Mantel test

4.5 Discussion

Understanding the response of fungal communities to elevated atmospheric CO_2 is important for global change biology. This study comprehensively surveyed soil fungal communities under aCO_2 and eCO_2 by 454 pyrosequencing of 28S rRNA gene amplicons and RMT-based ecological network analysis. Our results indicated several interesting mechanisms about how fungal communities respond to long-term eCO_2. First, long-term eCO_2 did not significantly change the overall fungal community

structure and species richness, but increased fungal diversity with higher evenness of overall abundance. Second, co-occurrence network analysis suggested that fungal communities respond to long-term eCO_2 by community reassembly. Third, such changed co-occurrence network topology was significantly correlated with increased soil ammonification rate and plant biomass, and OTUs belonging to Ascomycota were mainly responsible for such changes. These results will provide novel insights on how the ongoing global elevated atmospheric CO_2 affects the Earth's fungal community.

Our first hypothesis is that long-term eCO_2 would change the fungal community structure and diversity due to changed soil and plant properties. Unexpectedly, we did not see significant changes of overall fungal community structure and species richness between aCO_2 and eCO_2 samples, as revealed by both dissimilarity and NMDS ordination analysis. However, both taxonomic and phylogenetic diversity increased as a result of higher species evenness of overall abundance. Although no significant differences were observed at the phylum/order level, relative abundances of 119 OTUs (about 27% of all captured sequences) were significantly different between aCO_2 and eCO_2 fungal communities. Interestingly, decreased and increased overall relative abundances of Ascomycota and Basidiomycota in eCO_2 samples were observed for the 119 OTUs. Compared with Weber et al.'s recent study in a forest FACE site (Weber, Vilgalys et al. 2013), our results were generally consistent that eCO_2 had no significant effects on high level fungal groups when relative abundances for all OTUs were considered. Our results were also generally consistent with a previous study that the fungal richness was not significantly affected by eCO_2 (Parrent, Morris et al. 2006).

Another objective of this study is to determine the diversity and composition of fungal communities in the BioCON grassland ecosystem. The grassland soil ecosystem in the BioCON experimental site in Minnesota was dominated by Ascomycota (81% at aCO_2 and 77% at eCO_2) and Basidiomycota (11% at aCO_2 and 14% at eCO_2). Compared with the reports by previous studies (Buée, Reich et al. 2009; Jumpponen, Jones et al. 2010; Xu, Ravnskov et al. 2012; Penton, St. Louis et al. 2013; Weber, Vilgalys et al. 2013), fungal community composition in soil varied greatly across different types of soil ecosystems. Such variations in fungal community composition between different studies might be caused by different coverage of different primer sets or phylogenetic markers (such as ITS vs. 28S) (Toju, Tanabe et al. 2012), but more likely caused by plant species, soil, and/or climate differences (Buée, Reich et al. 2009). Nonetheless, all of these studies suggested Ascomycota and Basidiomycota be the dominant fungal phyla in soil ecosystems. Notably, only about 0.03% of reads (53 OTUs) were from Glomeromycota in this study. Glomeromycota is the phylum that most arbuscular mycorrhizal fungi belong to and was previously reported to be dominant in grasslands (Santos-González, Finlay et al. 2007) and widespread among different global ecosystems (ÖPik, Moora et al. 2006). Since a previous study using the same primer set identified at least 15% Glomeromycota in an Oklahoma tallgrass prairie soil, the low relative abundance of Glomeromycota identified in this study did not arise from the primer set used for PCR amplification, which was also verified by the NCBI primer tool (Penton, St. Louis et al. 2013). As arbuscular mycorrhizal fungi form symbioses with many herbaceous land plants, the low relative abundance of Glomeromycota may result from different plant species composition in these

ecosystems, as root hairs were not removed prior to DNA extraction and rhizosphere soil and bulk soil were not specifically distinguished during sampling process in either study.

Our second hypothesis is that long-term eCO_2 would promote fungal community reassembly, as driven by increased nutrition availability in the soil. To test this hypothesis, co-occurrence ecological network analysis was implemented. Ecological network analysis is a systems-level method to identify species interactions/co-occurrence within an ecosystem that cannot be directly observed (Fath, Scharler et al. 2007). Similar to the food web network analyses in macroecosystems, microorganisms including fungi should also form complex interactions with positive or negative impacts on other species (Faust and Raes 2012). *"It would not be surprising to see entire patterns of community organization jumbled as a result of global change"* (Kareiva, Kingsolver et al. 1993). For macroecosystems, many lines of evidence have shown that global change exerts pervasive impacts on various antagonistic and mutualistic interactions among species (Tylianakis, Didham et al. 2008). It is expected that the co-occurrence patterns of fungal communities would change in responding to changed soil and plant properties, by forming denser and more collaborative relationships in decomposing increased biomass in the soil. Comparative analysis of fungal co-occurrence networks in this study verified our hypothesis and indicated that long-term eCO_2 affected the fungal community in the following ways. First, eCO_2 increased the complexity of interactions within the fungal community, as evidenced by increased intermodular and intramodular connectivity, decreased geodesic distance, and decreased modularity, suggesting more intense interspecies correlations. Second, eCO_2 increased

negative relationships between fungal species/OTUs with lower relative abundances (all < 1%, except one), suggesting that increased carbon inputs into soil increased competition between less dominant fungal species. Third, eCO_2 decreased the connectivity for abundant OTUs. In the aCO_2 network, dominant OTUs formed relatively complex networks by co-occurring with other less abundant ones, while in the eCO_2 network, much fewer connections were observed for the same dominant OTUs. Finally, eCO_2 promoted fungal community reassembly. At least 31 OTUs that were sparsely distributed in different modules in the aCO_2 network became connected with each other and formed dense modules in the eCO_2 network, suggesting a possible community reassembly process.

Interestingly, the changed fungal network topology under eCO_2 was significantly correlated with increased plant biomass and NH_4^+ availability in the soil. This indicated that the increased plant biomass and NH_4^+ availability in the soil might be the driving force for the changed network topologies, providing novel insights into how fungal communities respond to eCO_2. Fungal communities are well known as decomposers in the ecosystem, by degrading organic matters into biologically available nutrients (Hu, Chapin et al. 2001; Chung, Zak et al. 2006; Cheng, Booker et al. 2012). Under eCO2, both aboveground and belowground plant biomass was stimulated (Reich, Knops et al. 2001; He, Xu et al. 2010; Langley and Megonigal 2010; Drake, Gallet-Budynek et al. 2011; Zak, Pregitzer et al. 2011; Reich and Hobbie 2013), providing more organic matters for fungal communities as well as proposing higher demand for biologically available nitrogen (Luo, Su et al. 2004; Luo, Hui et al. 2006; Reich, Hobbie et al. 2006; Reich and Hobbie 2013). In order to effectively degrade increased organic

matters in soil to provide more ammonium for stimulated plant growth, fungal species may tend to interact with each other more intensely, leading to reassembled community topology. Such increased network complexity was not only observed in fungal communities. Bacterial communities responded similarly to eCO_2, as revealed by both phylogenetic and functional microbial ecological networks (Zhou, Deng et al. 2010; Zhou, Deng et al. 2011).

In conclusion, our study suggested that microbial fungal communities mainly responded to long-term eCO_2 by community reassembly with the overall community structure and species richness unchanged. Such responses were closely related with altered soil and plant properties, especially with the plant biomass and NH_4^+ availability in soil, thus it is expected to sustain as long as the plant biomass is stimulated by eCO_2. However, previous studies have shown that the microbial decomposition and plant biomass stimulation by eCO_2 were constrained by N availability in natural soil ecosystems (Hu, Chapin et al. 2001; Luo, Su et al. 2004; Reich, Hobbie et al. 2006). Therefore, the described responses of fungal community to eCO_2 may subject to change when a new balance between microbial decomposition, plant biomass and N availability is reached.

Chapter 5: The Diversity and Co-occurrence Patterns of N_2-fixing Microorganisms in a CO_2 Enriched Grassland Ecosystem

5.1 Abstract

Diazotrophs are the major contributor responsible for atmospheric nitrogen (N_2) fixation into the Earth's biosphere. The extensive diversity and structure of N_2-fixing communities and their responses to increasing atmospheric CO_2 remain to be further explored. By pyrosequencing *of nifH* gene amplicons and extraction of *nifH* genes from shotgun metagenomes, coupled with co-occurrence ecological network analysis approaches, we comprehensively analyzed the diazotrophic community exposed to eCO_2 for 12 years. Long-term eCO_2 significantly increased the abundance of *nifH* genes, but did not change the overall *nifH* diversity and diazotrophic community structure. Taxonomic and phylogenetic analysis of *nifH* amplicons suggested a high diversity of *nifH* genes in the soil ecosystem, with the majority belonging to cluster I and II *nifH* genes. Co-occurrence ecological network analysis suggested a clear preference of co-occurrence patterns between diazotrophs and other microbial species, and different co-occurrence patterns were observed for different subgroups of diazotrophs, such as *Azospirillum*/Actinobacteria, *Mesorhizobium*/Conexibacter, and *Bradyrhizobium*/Acidobacteria. This indicated a potential attraction of these non-N_2-fixers by diazotrophs in the soil ecosystem. Interestingly, more complex co-occurrence patterns were found for free-living diazotrophs than commonly known symbiotic diazotrophs, which is consistent with the physical isolation nature of symbiotic diazotrophs from the environment by root nodules. The study provides novel insights of our understanding microbial ecology of soil diazotrophs in natural ecosystems.

Keywords: *nifH*; soil diazotrophs; community structure; co-occurrence patterns; elevated CO_2

5.2 Introduction

Biological nitrogen fixation (BNF), the reduction of atmospheric N_2 to biologically available ammonium, is the major pathway that atmospheric N_2 enters the Earth's biosphere and contributes about 128 Tg N per year in natural terrestrial ecosystems (Galloway, Dentener et al. 2004). BNF is catalyzed by diverse but limited groups of nitrogenase-containing bacteria and archaea known as diazotrophs. The nitrogenase enzyme is composed of two components—component I for N_2 reduction with two heterodimers encoded by *nifD* and *nifK*, and component II that couples ATP-hydrolyzing to interprotein electron transfer with two identical subunits encoded by *nifH* (Zehr, Jenkins et al. 2003). Among these, *nifH* encoding the nitrogenase reductase subunit has most sequences available and become a promising gene marker for analyzing nitrogen fixation of microbial communities in various environments (Zehr, Jenkins et al. 2003; Raymond, Siefert et al. 2004). Although phylogenetic analysis of taxonomically identified nitrogenase genes provided evidences of an ancient horizontal gene transfer of nitrogenase between archaea and bacteria, recent events of horizontal gene transfer were not observed (Gaby and Buckley 2014). The study also suggested an insignificant correlation of sequence dissimilarities between *nifH* and 16S rRNA genes at the species level, i.e. larger variation of dissimilarity was found among *nifH* genes (Gaby and Buckley 2014). However, the phylogenetic relationships derived from *nifH* genes agree well with those derived from 16S rRNA genes at higher levels (Young 1992; Zehr, Mellon et al. 1995), indicating *nifH* as a promising biological marker for N_2 fixation in microbial community analysis.

The *nifH* gene family has been widely used in many studies to analyze diazotrophic microbial communities in various environments, especially in marine and soil ecosystems (Izquierdo and Nüsslein 2006; Moisander, Shiue et al. 2006; Mohamed, Colman et al. 2008; Hsu and Buckley 2009; Zehr 2011; Groß kopf, Mohr et al. 2012; Wang, Quensen et al. 2013; Berthrong, Yeager et al. 2014; Collavino, Tripp et al. 2014). As a result, novel insights into the diversity and structure of N_2-fixing communities have been gained. In marine environments, cyanobacteria are generally regarded as the major microorganisms responsible for N_2 fixation and can be classified into three major groups, including filamentous non-heterocyst-forming *Trichodesmium*, filamentous heterocyst-forming symbionts, and single-celled or unicellular cyanobacteria (Zehr 2011). Although *Trichodesmium* has been assumed as the major N_2-fixing cyanobacteria (Zehr 2011), a recent study suggested that the contributions of other N_2 fixers were much more significant than previously estimated (Groszkopf, Mohr et al. 2012), indicating the important role that these less dominant N_2 fixers play. In soil, N_2 fixation is dominated by symbiotic bacteria that form root-nodule symbiotic relationships with plants (Cleveland, Townsend et al. 1999). Similar to that in ocean, it is believed that both symbiotic and free-living diazotrophs contribute significantly to the Earth's N budget (Cleveland, Townsend et al. 1999). Although many previous studies focused on the relationship between symbiotic diazotrophs and plants, the diversity and community structure of diazotrophic communities were analyzed by several studies recently (Izquierdo and Nüsslein 2006; Hsu and Buckley 2009; Gaby and Buckley 2011; Wang, Quensen et al. 2013; Berthrong, Yeager et al. 2014; Collavino, Tripp et al. 2014). It has also been pointed out that the N_2 fixation rate in soil

was significantly affected by diazotrophic community structure (Hsu and Buckley 2009). However, the extensive diversity and complex community structure in soil ecosystems remains to be further explored.

Natural ecosystems under increased atmospheric CO_2 concentration are subjected to progressive N limitation (Hu, Chapin et al. 2001; Luo, Su et al. 2004; Finzi, Moore et al. 2006; Tilman, Reich et al. 2006) due to the stimulated plant growth rate and limited biologically available N in soil. Such progressive N limitation not only constrains the sustainability of ecosystem responses to eCO_2 (Luo, Su et al. 2004; Tilman, Reich et al. 2006), but also suppresses the microbial decomposition rate in soil (Hu, Chapin et al. 2001). As biologically available N mainly comes from the microbial decomposition of biomass and BNF, the stimulated plant growth and suppressed microbial decomposition should have proposed higher demand for BNF in soil. Thus it is of crucial interest for ecologists and microbial ecologists to understand how the belowground diazotrophic microbial communities respond to eCO_2. A recent study of N_2-fixing bacteria communities in forest ecosystems suggested that N fertilization had a stronger effect on the diazotrophic community than eCO_2. However, the response of diazotrophic community diversity and structure in the grassland ecosystem, one of Earth's largest ecosystems, is still not clear yet, although a previous GeoChip survey suggested increased *nifH* abundance in this same BioCON experimental site (He, Xu et al. 2010; Xu, He et al. 2013).

Similar to that using a food web to describe the community structure of macroecosystems, the interactive relationships should also be considered when analyzing the community structure of microorganisms. Using ecological network

approaches, co-occurrence ecological networks of microbial communities would be constructed and analyzed (Zhou, Deng et al. 2010; Steele, Countway et al. 2011; Zhou, Deng et al. 2011; Barberan, Bates et al. 2012; Faust, Sathirapongsasuti et al. 2012). By implementing co-occurrence ecological network approaches, co-occurrence patterns can be identified for diazotrophic microorganisms, providing novel insights into how other microorganisms potentially interact with diazotrophs. Since symbiotic diazotrophs enter plant roots and form nodules that physically isolate them from the environment and are less likely to form complex relationships with free-living microorganisms, co-occurrence networks may also provide information to identify free-living diazotrophs from symbiotic diazotrophs.

In this study, by sequencing of *nifH* amplicons coupled with extraction of shotgun metagenome sequencing data and co-occurrence ecological network analysis, we aimed to reveal the response of soil diazotrophs to eCO_2, and to determine the diversity and structure of soil diazotrophs, as well as their co-occurrence patterns in the BioCON experimental site after 12-year exposure to eCO_2 (Reich, Knops et al. 2001). The following hypotheses were tested: (1) Increased plant growth would enhance the demand for biological N_2 fixation in soil, resulting in increased *nifH* gene abundance as well as changed diazotrophic community diversity and structure; and (2) Free-living rather than symbiotic diazotrophs would form more complex co-occurrence ecological networks, and different co-occurrence patterns would be observed for different diazotrophs. This study provides valuable insights into our understanding of microbial ecology of diazotrophs in soil.

5.3 Materials and Methods

5.3.1 Site description and sample collection

The same samples collected from the BioCON experimental site were used. Please refer to Chapter 3 for more detailes.

5.3.2 DNA extraction, purification and quantification

Soil DNA was extracted by freeze-grinding mechanical lysis as described previously (Zhou, Bruns et al. 1996), and was purified using a low melting agarose gel followed by phenol extraction for all 24 soil samples collected. DNA quality was assessed by the ratios of 260/280 nm, and 260/230 nm using a NanoDrop ND-1000 Spectrophotometer (NanoDrop Technologies Inc., Wilmington, DE), and final soil DNA concentrations were quantified with PicoGreen (Ahn, Costa et al. 1996) using a FLUOstar Optima (BMG Labtech, Jena, Germany).

5.3.3 PCR amplification and 454 pyrosequencing

A total of 23 samples instead of 24 were subjected to 454 pyrosequencing due to insufficient remaining DNA and soil for one of the samples. Amplification was performed using the *nifH* PolF/PolR primers (PoLF: TGCGAYCCSAARGCBGACTC, and PolR: ATSGCCATCATYTCRCCGGA), whose products are expected to be approximately 362-bp (Poly, Monrozier et al. 2001). A unique 8-mer barcode was added for each sample at the 5'-end of the forward primer. The barcode-primers were synthesized by Invitrogen (Carlsbad, CA) and used for the generation of PCR amplicons. Quadruplicate 20 µl PCR reactions were performed as follows: 4 µl Promega GoTaq buffer, 0.5 µl GoTaq DNA polymerase, 1.5 µl Roche 25 mM MgCl2, 1 µl Invitrogen 10 mM dNTP mix, 1 µl of each primer (10 pmol μl^{-1}), 0.2

μl New England BioLabs 10 mg ml^{-1}BSA, 1 μl 10 ng 7 μl^{-1} template, and 9.8 μl H_2O. Cycling conditions were an initial denaturation of 94 ℃ for 3 min, 30 cycles of 94 ℃ for 1 min, 51 ℃ for 40 s, 72 ℃ for 1 min, and an final extension at 72 ℃ for 10 min. PCR products were gel purified using the Qiagen Gel Purification Kit following band excision. Products were further purified using the Qiagen PCR purification kit. After adapter ligation, amplicons were sequenced on a FLX 454 system (454 Life Sciences, Branford, CT) by Macrogen (Seoul, South Korea) using Lib-L kits and processed using the shotgun protocol.

5.3.4 Data analysis

Raw pyrosequencing reads were extracted from the *sff* file using the *sffinfo* tool from Roche 454. Two files, a *fasta* file containing the sequence and a *qual* file containing the quality information, were generated and then converted into a *fastq* file using the python script "faqual2fastq2.py" that comes with the UPARSE pipeline (Edgar 2013). The quality filtering, chimera removal and OTU clustering were carried out using the UPARSE pipeline (Edgar 2013), which is a recently developed approach that identifies highly accurate OTUs from amplicon sequencing data. Only the reads with perfectly matched barcodes and maximum of 2 primer mismatches were kept for further analysis. Barcodes and primers were deleted from reads. The remaining reads were then truncated to 300 bp, and reads with expected errors >0.5 were discarded. The program FrameBot (Wang, Quensen et al. 2013) was used to correct potential frame shifts caused by sequencing errors and only reads whose translated proteins got mapped to reference *nifH* protein sequences with >30% identity were kept. The reads were then dereplicated, sorted, and clustered into candidate OTUs with an identity cutoff of 0.94,

which is the average nucleotide identity that approximately corresponds to the species cutoff of 16S rRNA genes (Konstantinidis and Tiedje 2005). Chimeric OTUs were then identified and removed by searching against the *nifH* reference sequences maintained and curated by Zehr et al. (http://pmc.ucsc.edu/~wwwzehr/research/database/) (Zehr, Jenkins et al. 2003). Finally, qualified reads were mapped to OTU reference sequences for relative abundance calculation.

Taxonomic assignment for *nifH* OTUs was carried out by searching OTU representative sequences against reference *nifH* sequences with known taxonomic information. A minimum recalculated global identity cutoff of 80% was used to filter BLAST results. A lowest common ancestor algorithm was applied for taxonomic assignment based on the best BLAST hits with highest global identity. Taxonomic information at genus level or higher was assigned. For phylogenetic analysis, representative OTU sequences were aligned by the MUSCLE program (Edgar 2004), and a phylogenetic tree was built by FASTTREE (Price, Dehal et al. 2009). Significance tests for different taxonomic groups and OTUs were performed by response ratio analysis (Hedges, Gurevitch et al. 1999) at 95% confidence interval level. UniFrac PCoA analysis was done by the FastUniFrac pipeline (Hamady, Lozupone et al. 2009). Species richness, evenness, and diversity indices were calculated by the Mothur package (Schloss, Westcott et al. 2009), with rarefaction analysis of 1000 bootstrap random sampling iterations and 0.1% incremental sampling efforts.

5.3.5 Co-occurrence ecological network construction

In order to identify co-occurrence relationships between diazotrophs and other microbial species, a 16S rRNA gene amplicon dataset from the same site was also

included. Relative abundance profiles were generated for both *nifH* (random subsampling of 1200 reads) and 16S rRNA (random subsampling of 18 000 reads) OTUs. Co-occurrence ecological networks were constructed and analyzed using the online MENA pipeline, which implements random matrix theory (RMT) for threshold identification (Deng, Jiang et al. 2012). The RMT approach identifies the threshold by observing a transition point of nearest-neighbor spacing distribution of eigenvalues from Gaussian to Poisson distribution, which are two universal extreme distributions (Zhou, Deng et al. 2010). The RMT-based approach is a reliable and robust tool for network construction and has been successfully applied to construct various networks, including gene regulatory networks (Luo, Yang et al. 2007; Yang, Harris et al. 2008; Zhou, He et al. 2010; Lin, Song et al. 2011; Lin, Ji et al. 2013), functional molecular ecological networks (Zhou, Deng et al. 2010), and phylogenetic molecular ecological networks (Zhou, Deng et al. 2011). To construct highly confident co-occurrence ecological networks, only OTUs presented in at least 10 samples were used for Pearson correlation coefficient calculations and zero was filled in for missing values for OTUs in paired samples. This made the correlation coefficient between two OTUs more statistically reliable. Ecological networks were then visualized by Cytoscape (Smoot, Ono et al. 2011).

5.4 Results

5.4.1 Effects of eCO_2 on plant biomass, soil N, and nifH gene abundance

The plant biomass (aboveground and root) and soil nitrogen levels (NO_3^- and NH_4^+) were collected and analyzed. Long-term elevated CO_2 significantly increased plant

biomass and ammonification rate in soil, but not nitrification (Fig. 5.1). Consistent with previous observations (Reich, Knops et al. 2001; He, Xu et al. 2010), significantly increased plant biomass was found for both aboveground ($P = 0.01$) and root ($P = 0.06$) biomass (Fig. 5.1A). Such increased aboveground plant biomass and root biomass would have imposed higher demand for biologically available N (NO_3^- and NH_4^+) in soil. To analyze the nitrification and ammonification rates in soil, NO_3^- and NH_4^+ concentrations were then measured using a semi-open core, one-month *in situ* incubation approach. No significant differences were observed for initial NO_3^- and NH_4^+ amount between aCO_2 and eCO_2 samples. After one-month *in situ* incubation, the NO_3^- amount in the soil increased from 0.53 mg/kg soil to 3.08 mg/kg soil (Fig. 5.1B). No significant differences were observed for the final NO_3^- amount between aCO_2 and eCO_2 samples, suggesting a similar nitrification rate of microbial communities under aCO_2 and eCO_2. Interestingly, the NH_4^+ amount in aCO_2 samples decreased significantly after incubation, while the amount in eCO_2 samples remained almost unchanged at ~4.3 mg/kg soil, resulting in significantly higher final NH_4^+ amount in eCO_2 samples (Fig. 5.1B). Since NO_3^- in natural ecosystems is mostly converted from NH_4^+, similar nitrification and higher ammonification in eCO_2 samples suggested that more nitrogen were fixed or converted from other sources, such as N_2 fixation by diazotrophs.

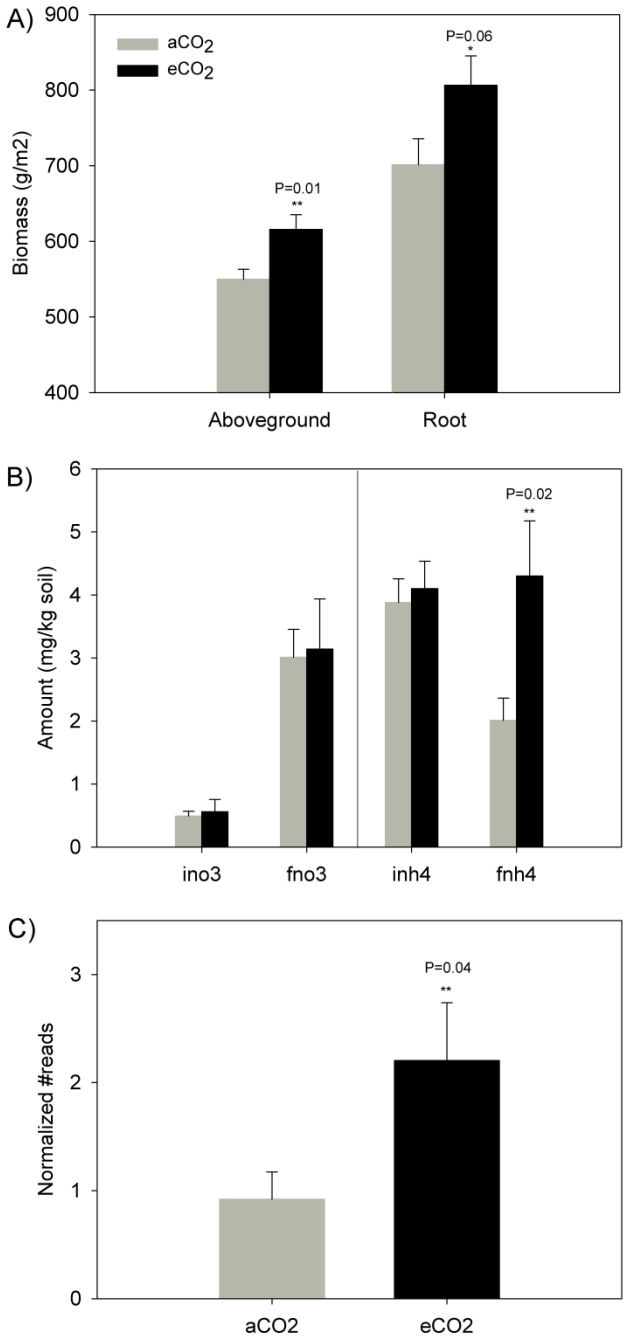


Fig. 5.1 Effects of long-term eCO_2 on plant biomass (A), soil NO_3^- and NH_4^+ (B), and *nifH* gene abundance (C). Both aboveground and root biomass were averaged from 5 years at the time of sampling, i.e. 2005-2009. Soil NO_3^- and NH_4^+ concentrations were then measured using a semi-open core, one-month *in situ* incubation approach. The abundance of *nifH* genes was obtained from shotgun metagenome datasets by extracting sequences mapped to *nifH* genes. Statistical testing was performed by the Student's t test. ino3 and fno3: initial and final NO_3^- concentration; inh4 and fnh4: initial and final NH_4^+ concentration. Grey bars represent aCO2 samples, and black bars for eCO2 samples.

The abundance of *nifH* genes was assessed by extracting *nifH* sequences from shotgun metagenome datasets, which were annotated by searching against eggNOG database (Muller, Szklarczyk et al. 2010). Comparisons were performed by randomly selecting 350,000 reads per sample from the shotgun metagenome. Although only a few *nifH* reads were identified, the *nifH* genes were twice abundant in eCO_2 samples than that in aCO_2 samples (Fig. 5.1C), with significance Student's t test p-value of 0.04. This suggested that eCO_2 had significantly increased the abundance of *nifH* genes in soil.

5.4.2 Sequencing data summary

Using 454 pyrosequenicng, a total of 102,679 raw forward reads targeting *nifH* gene amplicons were obtained for 23 samples with an average length of 338 bp. Four samples (two aCO_2 and two eCO_2) were excluded from further data analysis for their having < 500 reads. After quality trimming, frameshift correction and chimera removal, 73,161 reads were clustered into 749 *nifH* OTUs at 94% identity cutoff, of which 624 (a total of 73,036 reads with 42,725 from aCO_2 samples and 30,436 from eCO_2 samples) were non-singleton OTUs. The number of sequences in each sample ranged from 1,184 to 7,579 (3,851 on average), resulting in 80 to 287 OTUs per sample. A random re-sampling effort of 1200 reads per sample was made for further statistical analysis.

5.4.3 No significant eCO_2 effects on overall nifH-community diversity and structure

To analyze the *nifH*-community diversity in the grassland soil ecosystem and their responses to eCO_2, the OTU richness (Chao1), evenness, taxonomic and phylogenetic diversity indices were calculated (Fig. 5.2). A total of 633 and 616 OTUs were identified for aCO_2 and eCO_2 samples with the current sequencing effort. No significant differences between aCO_2 and eCO_2 samples were observed for the OTU richness, as

the 95% confidence intervals were overlapped with any number of randomly sampled sequences (Fig. 5.2A). Similarly, no significant differences were observed for the evenness of the overall *nifH* community between aCO_2 and eCO_2 sites (Fig. 5.2B), resulting in insignificant changes of the taxonomic diversity (Fig. 5.2C). Consistently, the phylogenetic diversity, which also considers the phylogenetic relationship among OTUs did not significantly change in response to eCO_2 (Fig. 5.2D). All these results suggested that the diversity of *nifH*-community was not significantly affected by long-term eCO_2 in the grassland ecosystem.

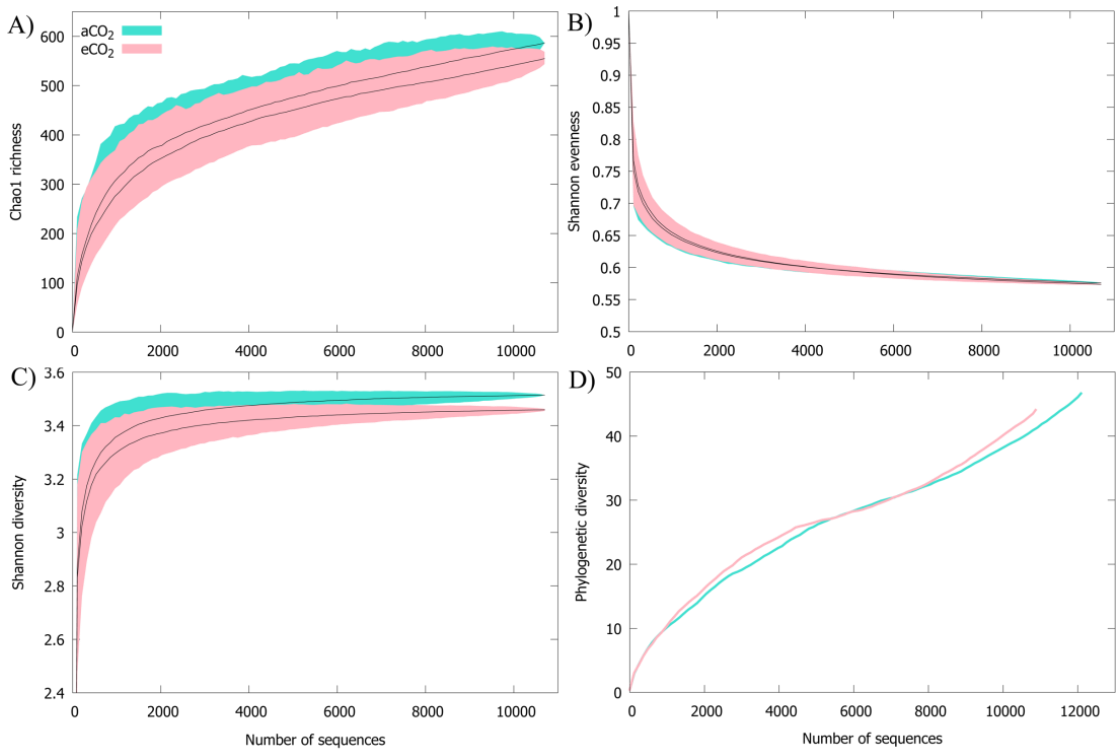


Fig. 5.2 The diversity of *nifH* genes in the grassland ecosystem under ambient CO_2 and elevated CO_2 conditions: (A) Chao1 richness; (B) Shannon evenness; (C) Shannon diversity; (D) phylogenetic diversity. Black line represents the averaged value for each diversity index. Turquoise and light-pink regions represent 95% confidence intervals.

Long-term eCO_2 did not significantly alter the overall *nifH*-community structure in the grassland soil ecosystem either (Fig. S5.1-S5.2). The overall community

structural differences among all samples were assessed by both weighted (with relative abundance data)/unweighted (with richness data) UniFrac PCoA analyses. A separation trend of eCO_2 samples from aCO_2 samples could be observed by unweighted UniFrac PCoA (Fig. S5.1), but not by weighted PCoA (Fig. S5.2). This indicated that the trend of separation of unweighted PCoA analysis should be due to the rare species rather than abundant ones. Further dissimilarity analysis also suggested that the overall community structure between aCO_2 and eCO_2 samples was not significantly different (ADONIS: $F = 0.062$, $P = 0.329$; ANOSIM: $R = 0.035$, $P = 0.249$; MRPP: $\delta = 0.531$, $P = 0.252$).

5.4.4 The taxonomic and phylogenetic composition of nifH genes

Unlike 16S rRNA genes, reference sequences for *nifH* genes from cultivated microbial strains/species are still very limited, making it difficult to classify *nifH* sequences into their taxonomic groups, especially at the species/strain level. We first tried a strict manner to only assign taxonomic information to OTUs having a minimum of 94% sequence identity with references in the *nifH* database. As a result, only 49 OTUs could be assigned to known taxonomic groups, among which six were assigned at 100% identity. Even at 90% sequence identity cutoff, this number only increased to 119, indicating a large diversified genetic pool of *nifH* gene variants in the soil microbial community. The taxonomic information for *nifH* OTUs was hence assigned as the lowest common ancestor of the best hits at a cutoff of 80% minimum sequence identity with reference sequences. Genus or higher taxonomic information was then assigned to 478 *nifH* OTUs as their nearest taxonomic matches.

The *nifH* community was dominated by Alphaproteobacteria as viewed by both OTU number and relative abundance, followed by Betaproteobacteria, Actinobacteria,

Delta-/Gamma-proteobacteria, and Bacilli (Fig. 5.3A). At the genus level, a total of 134 OTUs were assigned to *Bradyrhizobium* and accounted for 56.1% total sequences in the community. Other abundant genera detected with >1% relative abundance were *Mesorhizobium* (9 OTUs, 12.7% relative abundance), *Azospirillum* (20 OTUs, 4.8% relative abundance), *Azohydromonas* (4 OTUs, 3.3% relative abundance), *Frankia* (2 OTUs, 1.9% relative abundance), *Methylocystis* (12 OTUs, 1.45% relative abundance), and *Sideroxydans* (18 OTUs, 1.4% relative abundance). The most dominant OTU (OTU_1) belonged to *Bradyrhizobium*, and accounted for 35.3% of the *nifH*-containing community, followed by OTU_5 (*Mesorhizobium*, 8.34% relative abundance), OTU_2 (*Bradyrhizobium* , 5.36% relative abundance), and OTT_7 (*Azohydromonas*, 3.2% relative abundance). No significant changes of relative abundance were observed for the majority of OTUs although 34 OTUs did significantly change their abundances at eCO_2, which accounted for 13.15% of the total captured sequences. Among these, 18 were enriched in eCO_2 samples, and 16 were enriched in aCO_2 samples. All significantly changed OTUs were found with > 0.1% relative abundance in aCO_2 or eCO_2 samples (Fig. S5.3). Among these, five were found with >1% relative abundances, and were assigned to *Bradyrhizobium* (OUT_450), *Mesorhizobium* (OUT_711), *Azospirillum* (OUT_13), *Sideroxydans* (OUT_30), and *Frankia* (OUT_206) (Fig. S5.3). Notably, OTU_206 detected in 18 samples, which was also the only *Frankia* OTU found in more than three samples, increased significantly ($p < 0.05$) under eCO_2 (3.6% relative abundance in eCO_2 vs. 0.5% in aCO_2).

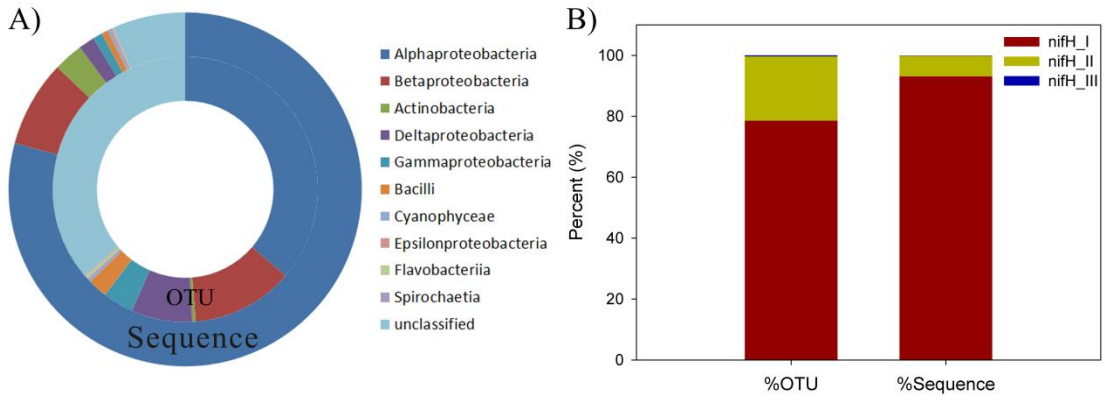


Fig. 5.3 Taxonomic (A) and phylogenetic (B) composition of *nifH* genes at both OTU and sequence levels. Taxonomic groups were summarized at the class level.

Phylogenetic clade assignment of *nifH* OTUs was performed at a lower cutoff (30% sequence identity) and the clade information of best BLAST hits aligned by each OTU was selected. The *nifH* community was dominated by sequences belonging to group I and group II NifH clades, which encode Mo-dependent nitrogenase (Zehr, Jenkins et al. 2003; Raymond, Siefert et al. 2004; Gaby and Buckley 2011), and accounted for 93.1% and 6.76% of the total captured sequences, and 78.54% and 21.13% total OTUs, respectively. Only two OTUs accounting for 0.14% of total sequences were found to be group III Mo-independent nitrogenase (Fig. 5.3B). No significant changes of relative abundances for any *nifH* groups were observed between aCO_2 and eCO_2.

5.4.5 Co-occurrence ecological networks of nifH communities

To explore the co-occurrence patterns of *nifH*-containing microorganisms with other microbial groups, co-occurrence ecological network was constructed using both 16S rRNA and *nifH* OTU profiles. By using the random matrix theory approach, a Pearson correlation coefficient cutoff of 0.81 was determined for network construction. Only the

first neighbors of *nifH* OTUs were extracted for further analysis, with the purpose to identify *nifH* OTU mediated co-occurrence patterns. As a result, six modules with more than 5 nodes were identified, covering 154 nodes and 242 links in total (Fig. S5.4).

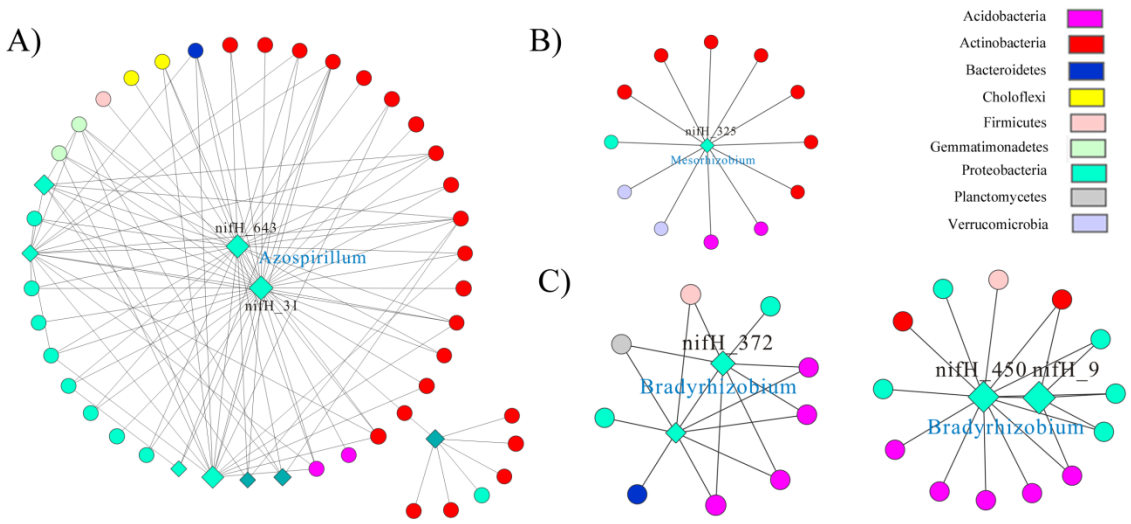


Fig. 5.4 Co-occurrence modules centered by *nifH* OTUs: (A) *Azospirillum* module; (B) *Mesorhizobium* module; (C) *Bradyrhizobium* module. *nifH* OTUs were represented by diamond shape. 16S rRNA OTUs were represented by circular shape. Different colors refer to different phyla.

Azospirillum module. The most complex module (module I, Fig. S5.4) was centered by two OTUs (*nifH*_643 and *nifH*_31) belonging to *Azospirillum*. Extraction of their first neighbors showed that these two *Azospirillum* OTUs were mainly connected by 17 Actinobacteria OTUs and 12 Proteobacteria OTUs, the latter of which included 4 *nifH* OTUs (Fig. 5.4A). Two Acidobacteria OTUs, two Chloroflexi OTUs, one Bacilli, one Sphingobacteria, and two unclassified *nifH* OTUs were also linked with the above *Azospirillum* OTUs. Among the 17 Actinobacteria OTUs connected with *Azospirillum* OTUs, nine were derived from Solirubrobacterales, six from Actinomycetales, and two from Acidimicrobiales. The connected Proteobacteria 16S rRNA OTUs were mainly

dominated by four Polyangiaceae OTUs, followed by two Rhizobiales, one Syntrophobacteraceae and one Oxalobacteraceae OTUs. These results suggested a high co-occurrence frequency of *Azospirillum* species with Actinobacteria species, especially those derived from Solirubrobacterales (Solirubrobacter and Conexibacter) and Actinomycetales.

Mesorhizobium module. Module II was centered by a *nifH* OTU (*nifH*_325) belonging to *Mesorhizobium* (Fig. 5.4B, Fig. S5.4). A total of seven Actinobacteria, two Acidobacteria, two Spartobacteria, and one Alphaproteobacteria OTUs were connected with the *Mesorhizobium* OTU. Notably, five of the Actinobacteria OTUs were assigned to Conexibacter, suggesting a high probability of co-occurrence relationship between *Mesorhizobium* and Conexibacter.

Bradyrhizobium modules. Two modules were centered by *Bradyrhizobium* (modules III, Fig. S5.4). In both modules, *Bradyrhizobium nifH* OTUs were connected by a high number of Acidobacteria species belonging to multiple subgroups, such as *Gp3*, *Gp4*, *Gp7* and *Gp17* (Fig. 5.4C). Specifically, the *nifH*_372 OTU was connected with three *Acidobacteria Gp4* species and one *Granulicella* species. Also, the *nifH*_450 OTU was connected with one *Acidobacteria Gp3*, two Gp4, one Gp7 and one Gp17 species. In addition, the *nifH*_450 OTU was connected with four Alphaproteobacteria 16S rRNA OTUs. These results suggested a high probability of co-occurrence patterns between *Bradyrhizobium* and Acidobacteria.

Modules IV and V. Module IV was a relatively simple module centered by two *Burkholderiales nifH* OTUs, which were connected with two Acidobacteria, three Proteobacteria, and one Actinobacteria species (Fig. S5.4). In contrast, module V was a

relatively complex module centered by one *Bradyrhizobium*, one *Desulfovibrio* and one unclassified *nifH* OTU (Fig. S5.4). The *Bradyrhizobium* OTU was connected with multiple 16S rRNA OTUs, without preferred co-occurrence patterns. However, the *Desulfovibrio* and unclassified *nifH* OTUs were connected with a high portion of Actinobacteria and Acidobacteria species.

5.6 Discussion

Understanding the diversity, composition and structure of N_2-fixing communities and their interactions with other groups is essential for reliably accessing and predicting N dynamics in ecosystems. In this study, we used next-generation sequencing and co-occurrence ecological network approaches to analyze N_2-fixing communities from grassland soils subjected to 12-year eCO_2 exposure. Our results showed that long-term eCO_2 significantly increased the abundance of *nifH* genes, but did not change the overall *nifH* diversity and structure of N_2-fixing communities. Co-occurrence ecological networks were observed with other microbial groups as well as within *nifH*-containing microorganisms. The study provides novel insights into our understanding microbial ecology of N_2-fixing communities in grassland ecosystems.

The first question is how soil *nifH* communities respond to long-term eCO_2 in this grassland ecosystem. As expected, long-term eCO_2 stimulated the plant growth rate, resulting in increased aboveground and belowground plant biomass, which is consistent with our previous observations (Reich, Knops et al. 2001; He, Xu et al. 2010; Reich and Hobbie 2013), as well as many other similar studies (Langley and Megonigal 2010; Drake, Gallet-Budynek et al. 2011; Zak, Pregitzer et al. 2011). Meta-analysis indicated that such increased plant growth rate as a result of eCO_2 imposed a higher demand for

biologically available N in soil, in the form of increased NH_4^+. Since NH_4^+ in natural ecosystems mainly originates from microbial fixation of atmospheric N_2 and decomposition of soil biomass, it is expected that key functional genes involved in microbial N_2 fixation would be changed. This was evidenced by the increased *nifH* gene abundance revealed by extraction of *nifH* genes from shotgun metagenome sequencing and our previous study using GeoChip that the abundance of *nifH* gene family and carbon degradation gene families increased as a result of eCO_2 (He, Xu et al. 2010; Xu, He et al. 2013). However, owing to the limited coverage of *nifH* probes on the array and low number of captured *nifH* sequences in shotgun metagenomes, the changes of *nifH* community structure and diversity in response to eCO_2 may not be reliably and comprehensively evaluated. To overcome such limitations, *nifH* amplicon sequencing approach was applied in this study. Strikingly, the overall *nifH* community structure and diversity did not change significantly in response to eCO_2. This suggested that long-term eCO_2 increased the overall *nifH* community abundance, but not necessarily changed the *nifH* diversity and community structure. Although such results contradicted our hypothesis that the imposed demand for more N by increased plant growth in response to eCO_2 would change the *nifH* diversity and community structure in the soil ecosystem, the observation was consistent with several recent studies (Law, Breitbarth et al. 2012; Berthrong, Yeager et al. 2014). For example, Berthrong et al. found relatively small effects of CO_2 treatment on N-fixing bacterial community in the four long-term eCO_2 experimental sites they investigated, and no consistent differences were observed for *nifH* diversity between aCO_2 and eCO_2 soils (Berthrong, Yeager et al. 2014). Such observation was also consistent with their another community level study

that eCO_2 induced shifts in microbial communities were driven by functional groups not necessarily possess *nifH* (Dunbar, Eichorst et al. 2012). Notably, although only two OTUs were found for the genus *Frankia*, they were about 8-fold abundant than that in aCO_2, which is also consistent with a previous study that the activity of *Frankia* species increased with infertile soil in response to eCO_2 (Koike, Izuta et al. 1997). Taken all our current and previous observations together, the long-term treatment of eCO_2 in this grassland ecosystem has increased the overall abundance of *nifH* gene family, but not necessarily changed the diazotrophic community diversity and structure.

Another objective in this study is to determine the diversity of *nifH* community in the grassland ecosystem. High diversity of *nifH* community was observed, with OTUs from six major phyla (ten classes), among which Proteobacteria (Alpha- and Beta-) is the most dominant groups, which is generally consistent with several previous studies in soil (Berthrong, Yeager et al. 2014; Collavino, Tripp et al. 2014). Although found with high diversity, 271 OTUs were still not classified to any taxonomic groups at 80% identity cutoff, suggesting a highly diverse genetic pool for *nifH* genes. Contrasting with Berthrong et al. and Collavino et al.'s study with *nifH* community in forest (Berthrong, Yeager et al. 2014) and pampas (Collavino, Tripp et al. 2014) soil ecosystem, low amounts of Deltaproteobacteria and cluster III/IV *nifH* genes were found in this study. Since the same PolF/PolR primer (Poly, Monrozier et al. 2001) was used for PCR amplification of the *nifH* genes, such differences should be due to the different plant composition in these ecosystems, which may favor distinct subsets of diazotroph communities (Mutch and Young 2004; Martinez-Romero 2009). Notably, *Bradyrhizobium* species, widely known as N_2-fixing bacteria forming symbiotic

relationships with legume species through nodules (Stacey 1995), are obviously the most abundant and dominant *nifH*-containing microorganisms in this soil ecosystem at both sequence and OTU levels, suggesting their major roles in N fixation in the grassland soil. More interestingly, the most abundant OTU, OUT_1, is 100% identical with *nifH* sequences from three *Bradyrhizobium* strains isolated from plant root nodules of *Centrosema virginianu, Centrosema virginianum* and *Lupinus perennis* collected in New York and North Carolina (Parker 2012), of which *Lupinus perennis* is also planted in this BioCON experimental site (Reich, Knops et al. 2001). This suggested a potential prevalent existence and plant-specific selection nature of some *Bradyrhizobium* species throughout the continent, though different geographical distance and soil properties they may inhabit.

The third interesting question we would like to address is that how *nifH*-containing microorganisms interact with others, i.e. who are the members included in the symbiotic N fixation niche they may form? Since the majority of microorganisms in soil are still uncultivable (Rappe and Giovannoni 2003), such co-occurrence relationships between microbial species can hardly be directly observed by currently available experimental procedure. In this study, we took advantage of amplicon sequencing of 16S rRNA genes and *nifH* genes, as well as random matrix theory based co-occurrence ecological network approach, to predict potential neighbors that co-occur with these *nifH*-containing microbial species. Such a rational design provided us opportunities to identify potential microbial interactions not only among *nifH*-containing microorganisms, but also between other microbial species and *nifH* communities. Although such approaches to identify *nifH* co-occurrence networks have

not previously performed, similar methods have been used to identify bacteria-diatom relationships (Stanish, O'Neill et al. 2013) and bacteria-archaea-protist relationships (Steele, Countway et al. 2011). As a result, several interesting messages have been brought to our attention. First, only a few OTUs belonging to *Bradyrhizobium* were included in the constructed co-occurrence ecological network, though *Bradyrhizobium* OTUs were most abundant at both OTU and sequence level. This is because most *Bradyrhizobium* species enter plant roots and form symbiotic relationships with plants in the form of root nodules (Stacey 1995), these species are physically isolated from other free living soil microorganisms, resulting in few co-occurrence patterns with other microbial species. Second, although not many, several *nifH* OTUs assigned to commonly known as symbiotic N_2-fixing bacteria such as *Mesorhizobium* and *Bradyrhizobium*, formed relatively complex co-occurrence ecological networks with other microbial species, confirming their potential role as free-living diazotrophs (Kahindi, Woomer et al. 1997; Okubo, Tsukui et al. 2012). Third, the taxonomic assignment of co-occurred 16S rRNA OTUs with *nifH* OTUs are distinctly different from these diazotrophs, contradicting a previous observation based on a global coexisting network using whole genome sequencing data that coexisting microorganisms are phylogenetically closely related and coexisting genomes tend to be more similar regarding pathway content and genome size (Chaffron, Rehrauer et al. 2010). This is possibly because the N_2-fixing ability of diazotrophs to produce NH_4^+ attracted more microbial species without such abilities, rather than other diazotrophs, which is also indirectly evidenced by a previous study that many bacteria prefer ammonia as nitrogen source (Müller, Walter et al. 2006).

Several co-occurrence modules centered by *nifH* OTUs were identified. Among them, the module centered by *Azospirillum* OTUs was the most complex one. Since *Azospirillum* species are usually isolated from the rhizosphere of various plant species and are free-living N-fixing bacteria closely associated with grasses (Tien, Gaskins et al. 1979; Reinhold, Hurek et al. 1987; Steenhoudt and Vanderleyden 2000; Eckert, Weber et al. 2001), the complex module formed by *Azospirillum* OTUs confirmed our hypothesis that free-living diazotrophs tend to form more complex networks than symbiotic ones. Clear preference of co-occurrence patterns were identified between diazotrophs and other microbial species, such as *Azospirillum*/Actinobacteria, *Mesorhizobium*/Conexibacter, and *Bradyrhizobium*/Acidobacteria, indicating different co-occurrence patterns for different diazotrophs. Such frequent co-occurrence patterns indicated a potential attraction of Actinobacteria and Acidobacteria by ammonium produced by potentially free-living diazotrophs. However, the exact underlying mechanism can hardly be identified with current approaches and knowledge.

In conclusion, this study comprehensively analyzed the diversity, structure and co-occurrence patterns of N_2-fixing microbial communities in a CO_2 enriched grassland ecosystem. Our results provided several valuable insights into the microbial ecology of N_2-fixing microorganisms and their responses to long-term eCO_2. First, this study was conducted in a grassland ecosystem subjected to >12 years eCO_2 treatment using multiple complementary approaches, providing reliable evidence that long-term eCO_2 affects microbial communities by increasing abundance of *nifH*-containing microorganisms. Second, the diversity and community structure identified in this study provided more information to our better understanding of the soil diazotrophs. Finally,

co-occurrence network analysis provided informative clues about how N_2-fixing microorganisms may interact with other species in the environment. Such information may also help to identify free-living N_2-fixers from symbiotic ones in a predictive manner. However, more experimental approaches are needed for accurate and reliable identification of species-species interactions.

Chapter 6: Summary and Output

How and by what mechanisms long-term eCO_2 affects belowground microbial communities is a critical issue for ecology and global change biology. By taking advantage of NGS technologies, this study comprehensively surveyed the response of microbial communities to long-term eCO_2 in a grassland ecosystem that had been exposed to CO_2 treatment for 12 years. Based on the observed results, several outcomes and/or mechanisms about microbial community responses to eCO_2 were revealed.

First, long-term eCO_2 decreased the belowground microbial biodiversity, including phylogenetic, taxonomic and functional diversity, as evidenced by 16S rRNA gene and shotgun metagenome sequencing. Until now, no consistent results were obtained about how elevated CO_2 affects the belowground microbial biodiversity and almost all previous studies about CO_2 effects on microbial diversity only focused on species richness (Janus, Angeloni et al. 2005; Lipson, Wilson et al. 2005; Lipson, Blair et al. 2006; Lesaulnier, Papamichail et al. 2008; Castro, Classen et al. 2010; Dunbar, Eichorst et al. 2012; Eisenhauer, Cesarz et al. 2012; He, Piceno et al. 2012). Our results provided comprehensive and solid evidences that elevated CO_2 decreases both taxonomic and functional richness and diversity, as well as phylogenetic diversity. Further investigation suggested that rare functional groups were mainly responsible for the decreased functional diversity, whereas diversity indices for major functional groups remained unchanged. Decreased taxonomic diversity was mainly contributed by Proteobacteria and less abundance taxonomic groups. Because the belowground microbial biodiversity is intimately linked to aboveground biodiversity (Wardle, Bardgett et al. 2004) and determines ecosystem multifunctioning (Fierer, Strickland et

al. 2009; Wagg, Bender et al. 2014), such decreased microbial biodiversity may lead to serious ecosystem consequences in the future.

Second, the decreased microbial biodiversity was a result of functional convergence to provide more biologically available N in the form of ammonium for stimulated plant growth. This is evidenced by decreased functional beta-diversity and increased taxonomic and phylogenetic beta-diversity of microbial communities under eCO_2, indicating that long-term eCO_2 selects microbial communities by function rather than taxonomy. Interestingly, the decreased microbial biodiversity was significantly correlated with increased ammonification rate in soil. Moreover, abundances for functional gene families that are responsible for producing ammonium from various sources were increased. Taken all evidence together, it could be concluded that long-term eCO_2 decreased microbial biodiversity by functional convergence, which is a process commonly found in macroecosystems (Reich, Walters et al. 1997; Paruelo, Jobbágy et al. 1998; Meinzer 2003; Shaver, Street et al. 2007), but not yet well established for microbial communities. This may be also a way that microbial communities mediate progressive N limitation (Hu, Chapin et al. 2001; Norby and Luo 2004; Reich, Hobbie et al. 2006; Norby, Warren et al. 2010; Reich and Hobbie 2013) as a result of continuously stimulated plant growth under eCO_2.

Third, the diversity and overall structure for fungal communities was not as significantly affected. Instead, fungal communities respond to long-term eCO_2 by community reassembly, which was significantly correlated with increased plant biomass and soil ammonification rate. No significant changes regarding the fungal community structure and species richness was observed, as revealed by NMDS ordination analysis

and diversity indices. However, the complexity of fungal community co-occurrence patterns significantly increased under eCO_2, as evidenced by increased inter-modular and intra-modular connections. Moreover, uncorrelated fungal species under aCO_2 formed highly dense co-occurrence modules under eCO_2, suggesting a process of community assembly. In contrast, the complexity of co-occurrences patterns for abundant fungal species decreased as a result of eCO_2. These changes of co-occurrence network topology were significantly correlated with increased plant biomass and soil ammonification rate, indicating community reassembly as a way for more efficient organic decomposition to produce more biologically available N for stimulated plant growth. The results provide novel insights into how fungal communities respond to long-term eCO_2 and increased plant biomass.

Fourth, long-term eCO_2 significantly increased the abundance of N fixation genes, but did not change the overall community structure and composition. Owing to the important roles that N fixation may play under eCO_2, under which progressive N limitation occurs due to stimulated plant growth, the diversity and community structure of N_2-fixing microbial community were analyzed. However, no significant changes of community structure and diversity were found. Instead, the abundance of N fixation gene—*nifH*, increased significantly, suggesting that eCO_2 affects the N_2-fixing community by increasing the overall abundance. In addition, the co-occurrence patterns for N_2-fixing community were analyzed. A clear preference of co-occurrence patterns between diazotrophs and other microbial species was found, and different co-occurrence patterns were observed for different subgroups of diazotrophs, such as *Azospirillum*/Actinobacteria, *Mesorhizobium*/Conexibacter, and

Bradyrhizobium/Acidobacteria. This indicated a potential attraction of these non-N_2-fixers by diazotrophs in soil ecosystems. Interestingly, more complex co-occurrence patterns were found for free-living diazotrophs than commonly known symbiotic diazotrophs, which is consistent with the physical isolation nature of symbiotic diazotrophs from the environment by root nodules. The study provides novel insights of our understanding microbial ecology of soil diazotrophs in natural ecosystems.

In addition, we also developed a new *k*-mer based computational approach that is able to identify microbial strains/species from complex shotgun metagenomes. This method can be well applied to analyze microbial communities with good coverage of reference genomes, such as human microbiomes. Sensitivity evaluation against synthetic metagenomes with different coverage suggested that 50 GSMs per strain were sufficient to identify most microbial strains with ≥ 0.25x coverage, and 10% of selected GSMs in a database should be detected for confident positive callings. We expect this method will be useful for microbial strain/species identifications in future soil metagenome studies with higher sequencing coverage. This approach is one of a few methods available for microbial identification at the strain level.

In conclusion, our study provided novel insights for better understanding the belowground microbial community and their responses to increasing atmospheric CO_2, and would be of great interest for microbiologists, ecologists, global change biologists, and bioinformaticians, and the developed novel data analysis methods and software tools are useful resources for the scientific community.

Those results from this study and other associated projects that I have involved are largely reflected in my publications (published, in press, in preparation) as they are listed below:

1. **Qichao Tu** et al. “Fungal communities respond to long-term elevated CO_2 by community reassembly”. (in revision)
2. **Qichao Tu** et al. “The diversity and co-occurrence patterns of N_2-fixing community in a CO_2 enriched grassland ecosystem”. (submitted)
3. **Qichao Tu** et al. “Long-term elevated CO_2 decreases microbial biodiversity by functional convergence”. (draft)
4. **Qichao Tu**, Zhili He, and Jizhong Zhou. "Strain/species identification in metagenomes using genome-specific markers". *Nucleic Acids Research* 42.8 (2014): e67-e67.
5. **Qichao Tu**, Zhili He, Yan Li, Yanfei Chen, Ye Deng, Lu Lin, Christopher L. Hemme, Tong Yuan, Joy Van Nostrand, Liyou Wu, Xuedong Zhou, Wenyuan Shi, Lanjuan Li, Jian Xu, Jizhong Zhou. “Development of HuMiChip for Functional Profiling of Human Microbiomes.” PLoS ONE 9(3): e90546. doi: 10.1371/journal.pone.0090546
6. **Qichao Tu**, Zhili He, Ye Deng and Jizhong Zhou. “Strain/Species-Specific Probe Design for Microbial Identification Microarrays.” *Applied and Environmental Microbiology*. *AEM.01124-13; doi:10.1128/AEM.01124-13*
7. **Qichao Tu**, Hao Yu, Zhili He, Ye Deng, Liyou Wu, Joy D. Van Nostrand, Aifen Zhou, James Voordeckers, Yong-Jin Lee, Yujia Qin, Christopher L. Hemme, Zhou Shi, Kai Xue, Tong Yuan, Aijie Wang, and Jizhong Zhou. “GeoChip 4: a functional gene array-based high throughput environmental technology for microbial community analysis." Molecular Ecology Resources. DOI:10.1111/1755-0998.12239 (in press)
8. **Qichao Tu**, Ye Deng, Jizhong Zhou, and Zhili He. “Development and Evaluation of Functional Gene Arrays with GeoChip as an Example” in *Microarrays: Current Technology, Innovations and Applications*. Caister Academic Press, Norwich, UK. ISBN: 978-1-908230-49-2 (Book Chapter, in press)
9. Chengwei Luo, Luis Rodriguez-R., Eric Johnston, Liyou Wu, Lei Cheng, Kai Xue, **Qichao Tu**, Ye Deng, Zhili He, Zhou Shi, mengting Yuan, Sherry Rebecca, dejun Li, Yiqi Luo, E.A.G. Schuur, Patrick Chain, James Tiedje, Jizhong Zhou, and Konstantinos Konstantinidis. “Soil microbial community responses to a decade of warming as revealed by comparative metagenomics”. *Applied and Environmental Microbiology* (2013): AEM-03712.

10. Meiying Xu, Yun Fang, Jun Liu, Xingjuan Chen, Guoping Sun, Jun Guo, Zhengshuang Hua, **Qichao Tu**, Liyou Wu, Jizhong Zhou, and Xueduan Liu. Draft genome sequence of *Shewanella decolorationis* S12, a dye degrading bacterium isolated from a waste-water treatment plant. *Genome announcements* 1.6 (2013): e00993-13.
11. Lu Lin, Yuetong Ji, **Qichao Tu**, Ranran Huang, Lin Teng, Xiaowei Zeng, Houhui Song, Kun Wang, Qian Zhou, Yifei Li, Qiu Cui, Zhili He, Jizhong Zhou, and Jian Xu. "Microevolution from shock to adaptation revealed strategies improving ethanol tolerance and production in thermophiles". *Biotechnology for Biofuels*. 2013, **6**:103 (doi:10.1186/1754-6834-6-103).
12. Fang YANG, Kang NING, Xingzhi CHANG, Xiao YUAN, Yue ZHANG, Xinping CUI, **Qichao TU**, Yuan TONG, Ye DENG, Christopher L Hemme, Joy Van Nostrand, Zhili HE, Jian Xu. "Saliva microbiota carry caries-specific functional gene signatures." *PLoS ONE* 9: e76458 (doi:10.1371/journal.pone.0076458).
13. Zhou, Aifen, He, Zhili, Qin, Yujia, Lu, Zhenmei, DENG, Ye, **Tu, Qichao**, Hemme, Christopher, Van Nostrand, Joy, Wu, Liyou, Hazen, Terry, Arkin, Adam, Zhou, Joe. "StressChip as a High Throughput Tool for Assessing Microbial Community Stability." *Environmental science & technology* 47.17 (2013): 9841-9849.
14. Dongru Qiu, Hehong Wei, **Qichao Tu**, Yunfeng Yang, Ming Xie, Jingrong Chen, Mark Pinkerton, Yili Liang, Zhili He, and Jizhong Zhou. "Combined genomics and experimental analyses of respiratory characteristics of Shewanella putrefaciens W3-18-1." *Applied and environmental microbiology* 79.17 (2013): 5250-5257.
15. Lee, Y.-J., J. D. Van Nostrand, **Q. Tu**, T. Yuan, L. Cheng, Z. Lu, Y. Deng, M. Q. Carter, Z. He, L. Wu, F. Yang, J. Xu, and J. Zhou. The PathoChip, a functional gene array for assessing pathogenic properties of diverse microbial communities. *The ISME journal*, *7*(10), 1974-1984.
16. Zhou, Jizhong, Liyou Wu, Ye Deng, Xiaoyang Zhi, Yi-Huei Jiang, **Qichao Tu**, Jianping Xie, Joy D. Van Nostrand, Zhili He, and Yunfeng Yang. "Reproducibility and quantitation of amplicon sequencing-based detection." *The ISME Journal* 5, no. 8 (2011): 1303-1313.
17. Hemme, Christopher L., Matthew W. Fields, Qiang He, Ye Deng, Lu Lin, **Qichao Tu**, Housna Mouttaki et al. "Correlation of genomic and physiological traits of thermoanaerobacter species with biofuel yields." *Applied and Environmental Microbiology* 77, no. 22 (2011): 7998-8008.
18. Lin, Lu, Houhui Song, **Qichao Tu**, Yujia Qin, Aifen Zhou, Wenbin Liu, Zhili He, Jizhong Zhou, and Jian Xu. "The thermoanaerobacter glycobiome reveals mechanisms of pentose and hexose co-utilization in bacteria." *PLoS Genetics* 7, no. 10 (2011): e1002318.
19. Zhou, Jizhong, Ye Deng, Feng Luo, Zhili He, **Qichao Tu**, and Xiaoyang Zhi. "Functional molecular ecological networks." *MBio* 1, no. 4 (2010).

20. He, Zhili, Ye Deng, Joy D. Van Nostrand, **Qichao Tu**, Meiying Xu, Christopher L. Hemme, Xingyuan Li et al. "GeoChip 3.0 as a high-throughput tool for analyzing microbial community composition, structure and functional activity." *The ISME Journal* 4, no. 9 (2010): 1167-1179.

Appendix A: Supplementary Tables

Table S2.1 Summary of mapped reads and GSMs of mock community metagenomes. Red denote false positives.

Table S3.1. Null model analysis of the centroids of soil microbial communities under aCO_2 and eCO_2 conditions.

Table S3.2 Alpha diversity at functional category levels.

Table S3.3 Spearman's correlation analysis between plant biomass and microbial biodiversity indices.

Table S4.1 Topological property of co-occurrence networks of fungal communities under eCO_2 and eCO_2 conditions and their.

Table S2.1 Summary of mapped reads and GSMs of mock community metagenomes. Red denotes false positives.

Accession	Strain Name	#reads mapped to genome	#reads mapped to GSMs	#mapped GSMs
SRR172902 (Even mock community, Illumina 75bp)	Streptococcus agalactiae 2603V/R	6104	1	5
	Streptococcus mutans UA159	113260	38	43
	Bacillus cereus ATCC 10987	52022	2	5
	Actinomyces odontolyticus ATCC 17982	154918	65	37
	Bacteroides vulgatus ATCC 8482	607992	98	46
	Acinetobacter baumannii ATCC 17978	787612	79	48
	Clostridium beijerinckii NCIMB 8052	243616	32	28
	Deinococcus radiodurans R1	2492460	802	50
	Enterococcus faecalis OG1RF	83409	30	21
	Lactobacillus gasseri ATCC 33323	1771	2	3
	Listeria monocytogenes EGD-e	123554	34	28
	Methanobrevibacter smithii ATCC 35061	44131	22	23
	Pseudomonas aeruginosa PAO1	47377	6	17
	Rhodobacter sphaeroides 2.4.1	211977	68	39
	Staphylococcus aureus subsp. aureus USA300_TCH959	209230	19	21
	Staphylococcus epidermidis ATCC 12228	330608	423	48
SRR072233 (Even mock community, 454 shotgun)	Streptococcus agalactiae 2603V/R	1227	3	4
	Streptococcus mutans UA159	22469	18	36
	Bacillus cereus ATCC 10987	10752	9	12
	Actinomyces odontolyticus ATCC 17982	34654	214	48
	Bacteroides vulgatus ATCC 8482	122608	266	50
	Acinetobacter baumannii ATCC 17978	174655	269	50
	Clostridium beijerinckii NCIMB 8052	45573	99	39
	Deinococcus radiodurans R1	524593	2511	47
	Enterococcus faecalis OG1RF	17847	63	37
	Lactobacillus gasseri ATCC 33323	391	2	2
	Listeria monocytogenes EGD-e	28108	86	45
	Methanobrevibacter smithii ATCC 35061	8128	54	30
	Pseudomonas aeruginosa PAO1	10460	15	20
	Rhodobacter sphaeroides 2.4.1	34127	186	47
	Staphylococcus aureus subsp. aureus USA300_TCH959	39582	60	27
	Staphylococcus epidermidis ATCC 12228	68187	746	49
	Streptococcus agalactiae 2603V/R	38388	6	15
	Streptococcus mutans UA159	625898	134	50
	Bacillus cereus ATCC 10987	24505	1	4
	Actinomyces odontolyticus ATCC 17982	786	0	0

SRR172903 (Staggered mock community, Illumina 75bp)	Bacteroides vulgatus ATCC 8482	1858	0	0
	Acinetobacter baumannii ATCC 17978	35169	1	1
	Clostridium beijerinckii NCIMB 8052	97505	16	16
	Deinococcus radiodurans R1	22254	4	5
	Enterococcus faecalis OG1RF	899	0	0
	Lactobacillus gasseri ATCC 33323	582	0	0
	Listeria monocytogenes EGD-e	6285	1	1
	Methanobrevibacter smithii ATCC 35061	307412	111	49
	Pseudomonas aeruginosa PAO1	200837	7	11
	Rhodobacter sphaeroides 2.4.1	2069571	387	50
	Staphylococcus aureus subsp. aureus USA300_TCH959	1838375	122	28
	Staphylococcus epidermidis ATCC 12228	1732754	1836	50
	Streptococcus pneumoniae SP3-BS71		1	7
	Mycobacterium tuberculosis H37Ra		1	2
	Staphylococcus aureus subsp. aureus TCH130		1	4
SRR072232 (Staggered mock community, 454 shotgun)	Streptococcus agalactiae 2603V/R	6712	5	16
	Streptococcus mutans UA159	106100	138	50
	Bacillus cereus ATCC 10987	4146	3	5
	Actinomyces odontolyticus ATCC 17982	138	1	3
	Bacteroides vulgatus ATCC 8482	324	1	2
	Acinetobacter baumannii ATCC 17978	6089	14	35
	Clostridium beijerinckii NCIMB 8052	18661	40	22
	Deinococcus radiodurans R1	2999	20	15
	Enterococcus faecalis OG1RF	132	0	0
	Lactobacillus gasseri ATCC 33323	106	0	0
	Listeria monocytogenes EGD-e	1009	4	5
	Methanobrevibacter smithii ATCC 35061	53719	331	49
	Pseudomonas aeruginosa PAO1	29446	61	50
	Rhodobacter sphaeroides 2.4.1	213549	1327	50
	Staphylococcus aureus subsp. aureus USA300_TCH959	309205	362	28
	Staphylococcus epidermidis ATCC 12228	310448	4114	50

Table S3.1. Null model analysis of the centroids of soil microbial communities under aCO_2 and eCO_2 conditions

	aCO2				eCO2			
	Centroid of actual communities	Centroid of null models	F	P	Centroid of actual communities	Centroid of null models	F	P
Taxonomic beta diversity (16S)	0.322	0.522	791.9	9.6e-19	0.333	0.526	352.5	3.6e-14
Taxonomic beta diversity (shotgun)	0.06	0.603	3292.5	1.9e-25	0.072	0.6	3484.9	9.9e-26
Functional beta diversity	0.098	0.605	3306.8	1.8e-25	0.092	0.602	13026.2	5.3e-32

Table S3.2 Alpha diversity at functional category levels

Functional Category	Chao1 richness			Shannon evenness			Shannon diversity		
	aCO_2	eCO_2	P	aCO_2	eCO_2	P	aCO_2	eCO_2	P
Function unknown	8230±258	7811±236	0.24	0.88±0.0011	0.88±0.0008	0.70	7.58±0.02	7.54±0.02	0.18
General function prediction only	**2603±87**	**2385±77**	**0.07**	0.72±0.0016	0.72±0.0013	0.15	5.36±0.02	5.34±0.01	0.34
Signal transduction mechanisms	1006±151	810±51	0.24	0.59±0.0036	0.59±0.0025	0.71	3.65±0.07	3.56±0.01	0.25
Amino acid transport and metabolism	967±25	967±36	0.99	0.77±0.0024	0.77±0.0016	0.14	5.00±0.01	5.00±0.01	0.55
Carbohydrate transport and metabolism	875±37	811±31	0.20	0.74±0.002	0.73±0.0014	0.81	4.70±0.03	4.66±0.02	0.22
Posttranslational modification, protein turnover, chaperones	865±97	766±56	0.39	0.74±0.0029	0.75±0.0025	0.44	4.62±0.06	4.53±0.02	0.19
Energy production and conversion	802±33	763±26	0.37	0.74±0.0013	0.74±0.0012	0.96	4.70±0.02	4.66±0.01	**0.08**
Transcription	788±94	673±36	0.28	0.67±0.0028	0.68±0.0021	0.24	4.14±0.03	4.10±0.01	0.27
Replication, recombination and repair	602±36	559±26	0.34	0.75±0.003	0.75±0.0017	0.23	4.52±0.01	4.48±0.01	**0.02**
Lipid transport and metabolism	573±37	541±25	0.49	0.62±0.0028	0.62±0.0019	0.27	3.66±0.03	3.63±0.01	0.25
Cell wall/membrane/envelope biogenesis	528±17	494±12	0.11	0.75±0.0016	0.75±0.0014	**0.00**	4.45±0.01	4.46±0.01	0.33
Inorganic ion transport and metabolism	517±28	489±20	0.42	0.71±0.0018	0.71±0.0016	0.22	4.21±0.03	4.19±0.01	0.53
Secondary metabolites biosynthesis, transport and catabolism	**466±13**	**410±13**	**0.01**	0.53±0.0036	0.53±0.0031	0.40	3.00±0.03	2.94±0.02	**0.08**
Translation, ribosomal structure and biogenesis	463±51	422±27	0.48	0.84±0.0047	0.84±0.0027	0.49	4.71±0.03	4.67±0.01	0.29
Coenzyme transport and metabolism	**407±9**	**373±10**	**0.02**	0.82±0.0016	0.82±0.0017	**0.09**	4.64±0.01	4.63±0.01	0.31
Intracellular trafficking, secretion, and vesicular transport	372±69	293±28	0.31	0.75±0.0047	0.75±0.0038	0.37	3.95±0.09	3.85±0.03	0.28
Nucleotide transport and metabolism	299±18	287±14	0.59	0.80±0.003	0.81±0.0017	0.40	4.25±0.02	4.21±0.01	0.13
Cell cycle control, cell division, chromosome partitioning	195±30	163±21	0.40	0.69±0.0056	0.68±0.004	0.68	3.09±0.09	2.99±0.03	0.30
Defense mechanisms	190±16	162±10	0.15	0.55±0.0022	0.55±0.0023	0.84	2.60±0.01	2.58±0.01	0.33
Cytoskeleton	167±49	111±15	0.29	0.57±0.029	0.54±0.0274	0.41	2.28±0.23	1.99±0.16	0.31
Cell motility	103±4	102±5	0.88	0.75±0.0029	0.75±0.0027	0.25	3.29±0.01	3.30±0.02	0.70
Chromatin structure and dynamics	89±19	64±11	0.27	0.60±0.0305	0.56±0.0217	0.33	2.11±0.22	1.87±0.12	0.33

Table S3.3 Spearman's correlation analysis between plant biomass and microbial biodiversity indices

	Aboveground biomass		Root biomass		Total biomass	
	Spearman's rho	P	Spearman's rho	P	Spearman's rho	P
Phylogenetic diversity	0.05	0.82	0.21	0.34	0.22	0.31
Taxonomic richness	0.17	0.43	-0.08	0.71	-0.09	0.68
Taxonomic diversity	0.19	0.39	-0.1	0.65	-0.06	0.8
Functional richness	-0.03	0.9	-0.25	0.23	-0.18	0.40
Functional diversity	-0.09	0.68	-0.3	0.15	-0.25	0.24

Table S4.1 Topological property of co-occurrence networks of fungal communities under eCO_2 and eCO_2 conditions and their corresponding random networks

	Experimental networks								Random networks		
	Total nodes	Total links	Neg. links	Avg. connectivity	Avg. geodesic distance	Avg. clustering coefficient	#Modules (≥3 nodes)	Modularity	Avg. geodesic distance	Avg. clustering coefficient	Avg. modularity
aCO2	271	647	37	4.78	6.0*	0.413	19 (12)	0.86*	3.53±0.09	0.028±0.006	0.43±0.007
eCO2	226	600	47	5.31	5.34*	0.391	13 (9)	0.80*	3.33±0.07	0.039±0.006	0.4±0.008

*Significant difference ($P<0.001$) between aCO_2 and eCO_2 fungal networks

Appendix B: Supplementary Figures

Fig. S2.1 Number of candidate GSMs when different k-mer sizes were used for continuous stretch filtering.

Fig. S2.2 Distribution of mapped GSM numbers to simulated metagenomes at sequencing coverage of 0.25, 0.5 and 0.75 with 50 and 100 GSMs/strain used.

Fig. S2.3 Comparison with MetaPhlAn at species level using synthetic metagenomes generated from 302 recently sequenced microbial genomes.

Fig. S3.1 Soil nitrification (A) and net N mineralization rate (B) in aCO_2 and eCO_2 samples summarized at two time frames: years 1-4 and years 5-12. Nitrification rate was suppressed by eCO_2 in years 1-4, but restored in years 5-12 ($P<0.1$).

Fig. S3.2 The Chao1 richness of taxonomic (A-D), genetic (E-F) and functional groups in aCO_2 and eCO_2 samples. Both alpha (left) and gamma (right) level richness was analyzed.

Fig. S3.3 The Shannon evenness of taxonomic (A-D), genetic (E-F) and functional groups in aCO_2 and eCO_2 samples. Both alpha (left) and gamma (right) level evenness was analyzed.

Fig. S3.4 The Shannon diversity for describing genetic and taxonomic diversities identified from shotgun metagenomes. Alpha (left), beta (middle) and gamma (right) diversities were calculated. **: $P<0.05$.

Fig. S3.5 The Chao1 richness (A,B), Shannon evenness (C,D) and Shannon diversity (E,F) for COG (left) and NOG (right) orthologous groups.

Fig. S3.6 The relative abundance of major bacteria phyla determined by 16S and shotgun metagenomes.

Fig. S3.7 The Chao1 richness (A-C), Shannon evenness (D-F), and Shannon diversity (G-I) for Actinobacteria, Proteobacteria and other rare phyla.

Fig. S3.8 Soil and plant properties under aCO_2 and eCO_2 collected in year 2009. (A) soil moisture, (B) pH, (C) nitrification rate, (D) ammonification, (E) aboveground biomass and (F) root biomass.

Fig. S3.9 Response ratio analysis of gene families related with NH_4^+ (left). Their relative abundances (middle) and major roles (right) in ammonium pathways are also plotted. Red indicate increased relative abundance, and green indicated decreased relative abundance.

Fig. S3.10 Response ratio analysis of gene families involved in nitrification and assimilatory NO_3^- reduction.

Fig. S4.1 eCO2 effects on soil moisture (A) and soil pH (B). No significant changes were observed for both proportional soil moisture and pH.

Fig. S4.2 Nonmetric multidimensional scaling analysis of overall fungal community structure under aCO_2 and eCO_2 samples. No clear separations could be observed.

Fig. S4.3 Rarefaction analysis of fungal community species richness (A), phylogenetic diversity (B), species evenness (C), and taxonomic diversity (D) under aCO2 and eCO2 samples. Filled curves refer to 95% confidence intervals.

Fig. S4.4 The composition of fungal community at (A) phylum level and (B) order level. Only the top 15 most abundant fungal orders with relative abundance $\geq 0.8\%$ were displayed. These 15 fungal orders accounted for about 75% of the total captured fungal community. Calculation was based on total number of sequences covered by OTUs.

Fig. S4.5 An overview of constructed networks for fungal communities at aCO_2 (A) and eCO_2 (B). More intense connections between nodes and modules were observed in eCO_2 network, showing more complex community interactions under eCO_2. Each node represents an OTU, which could be regarded as a fungal species. The size of nodes

represents relative abundance of OTUs. Each link connects two OTUs. Grey links means positive connections, and red means negative connections. Different colors refer to different fungal phyla.

Fig. S5.1 Unweighted UniFrac PCoA analysis of the *nifH* community. A trend of separation was found at all three dimensions analyzed.

Fig. S5.2 Weighted UniFrac PCoA analysis of the *nifH* community. The trend of separation disappeared when relative abundance of *nifH* OTUs was considered.

Fig. S5.3 Response ratio analysis of significantly changed *nifH* OTUs. Relative abundance and genus assignment for these OTUs were also included. Error bars plotted at the right side of the dashed line indicate significantly increased relative abundance at eCO_2, while error bars plotted at the left side indicate significantly decreased relative abundance at eCO_2.

Fig. S5.4 All *nifH*-centered modules identified in this study. Only modules with >5 nodes were included.

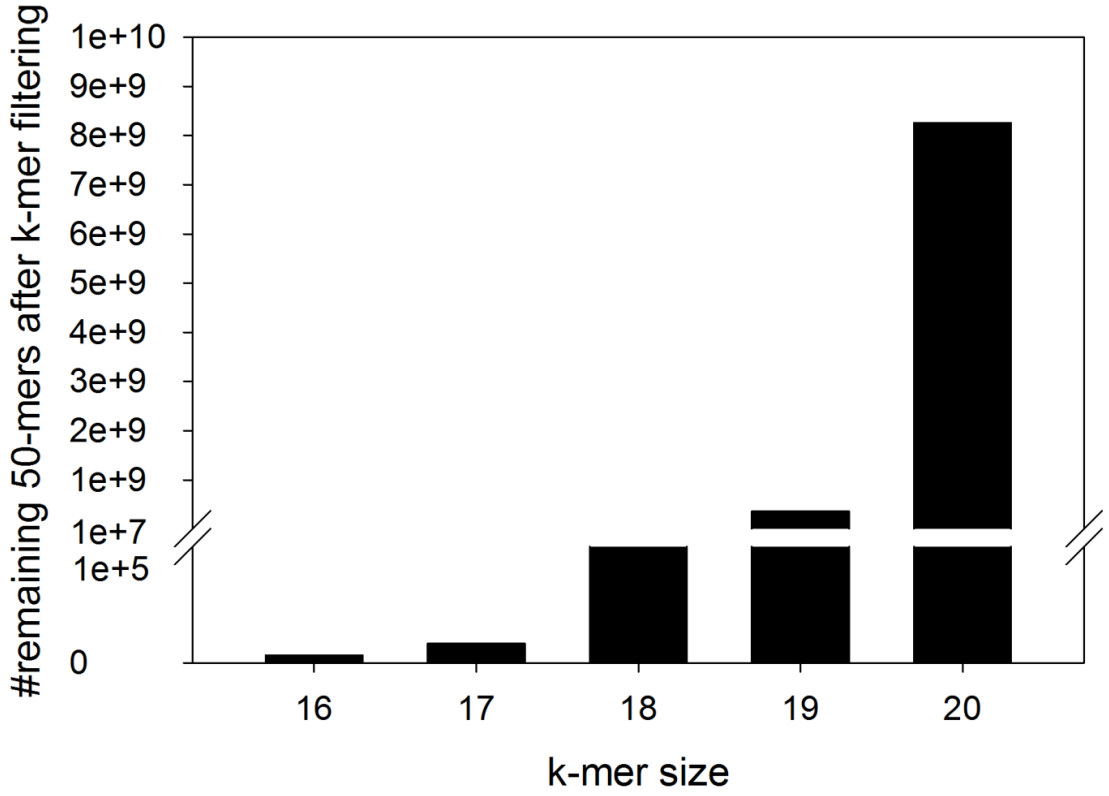


Fig. S2.1 Number of candidate GSMs when different k-mer sizes were used for continuous stretch filtering.

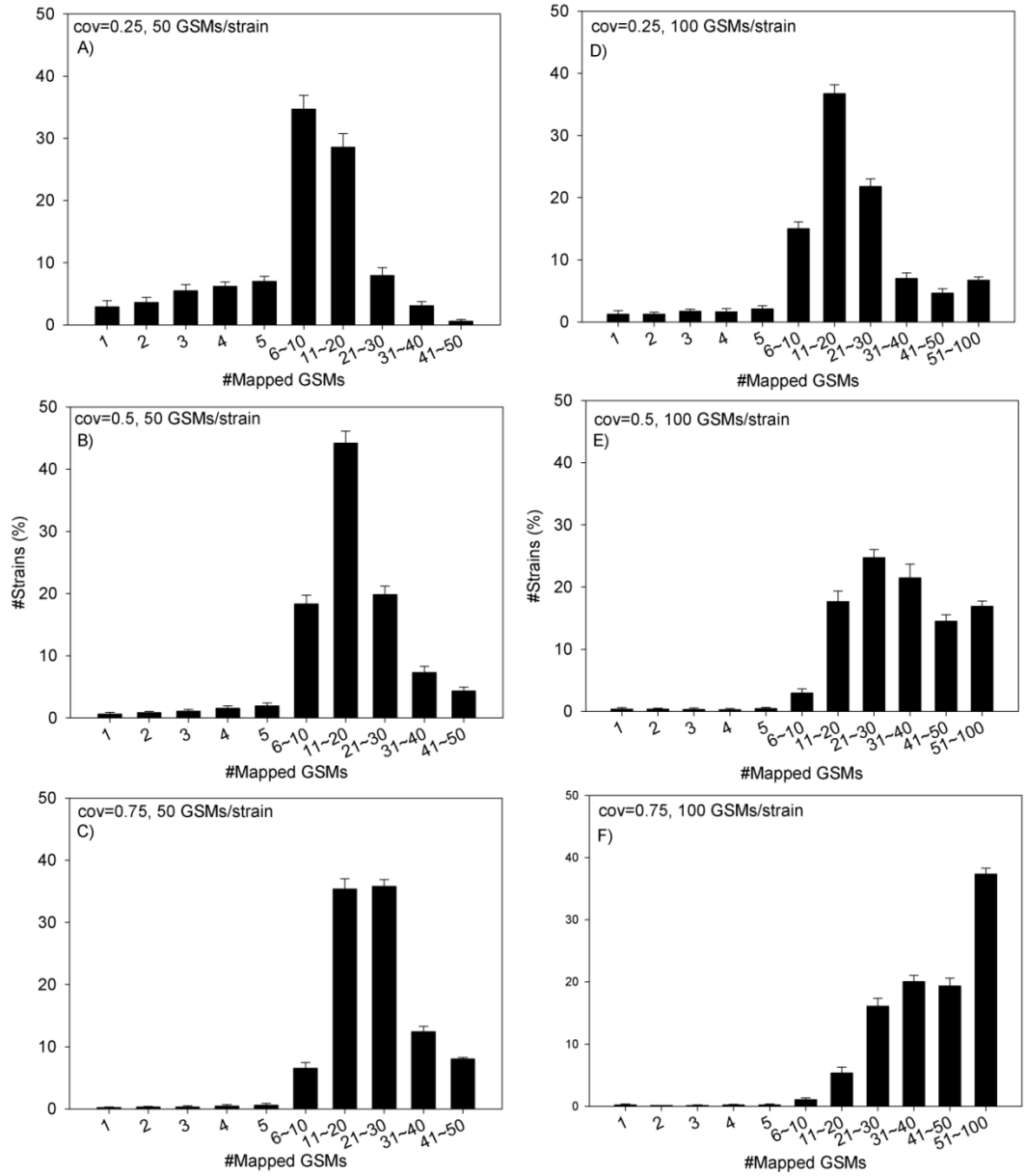


Fig. S2.2 Distribution of mapped GSM numbers to simulated metagenomes at sequencing coverage of 0.25, 0.5 and 0.75 with 50 and 100 GSMs/strain used.

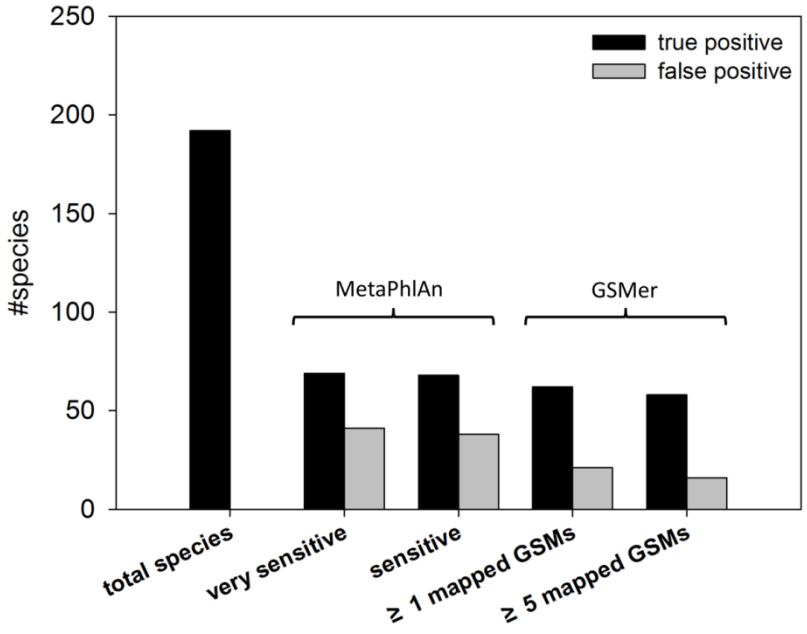


Fig. S2.3 Comparison with MetaPhlAn at species level using synthetic metagenomes generated from 302 recently sequenced microbial genomes.

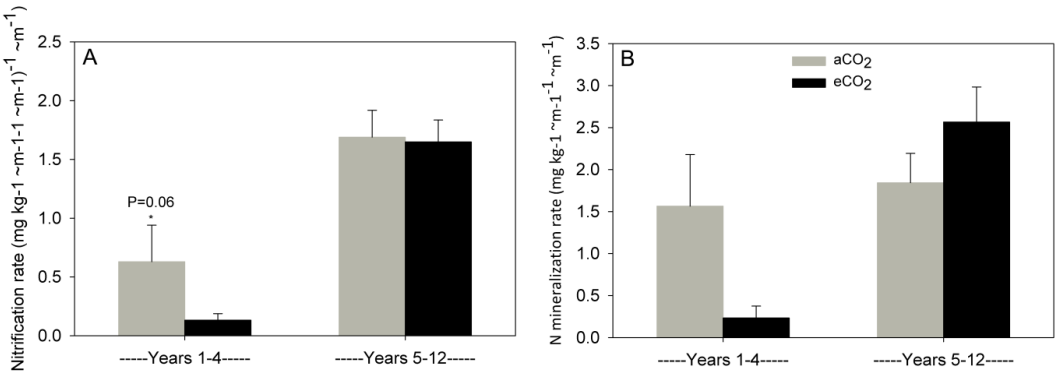


Fig. S3.1 Soil nitrification (A) and net N mineralization rate (B) in aCO_2 and eCO_2 samples summarized at two time frames: years 1-4 and years 5-12. Nitrification rate was suppressed by eCO_2 in years 1-4, but restored in years 5-12 ($P<0.1$).

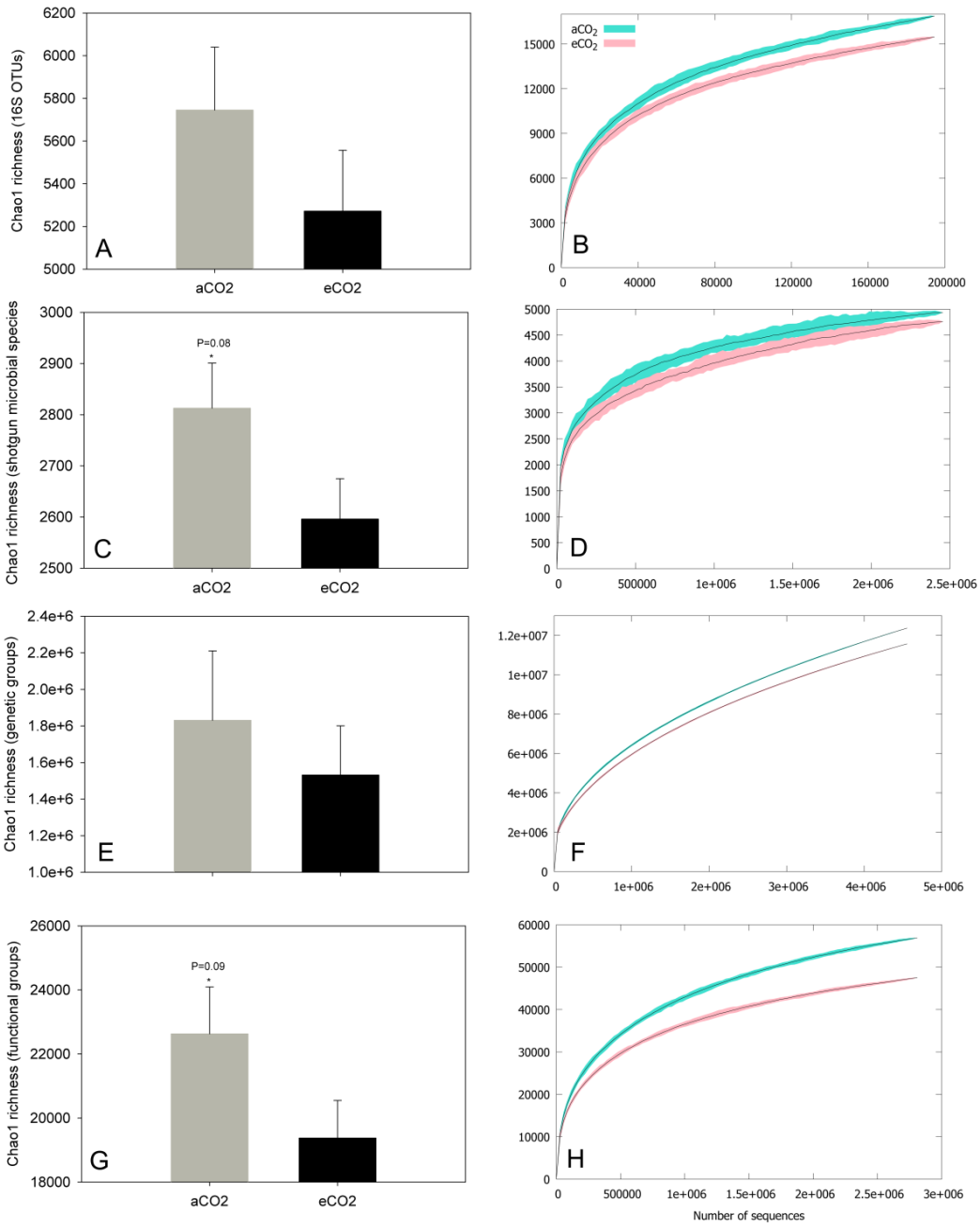


Fig. S3.2 The Chao1 richness of taxonomic (A-D), genetic (E-F) and functional groups in aCO_2 and eCO_2 samples. Both alpha (left) and gamma (right) level richness was analyzed.

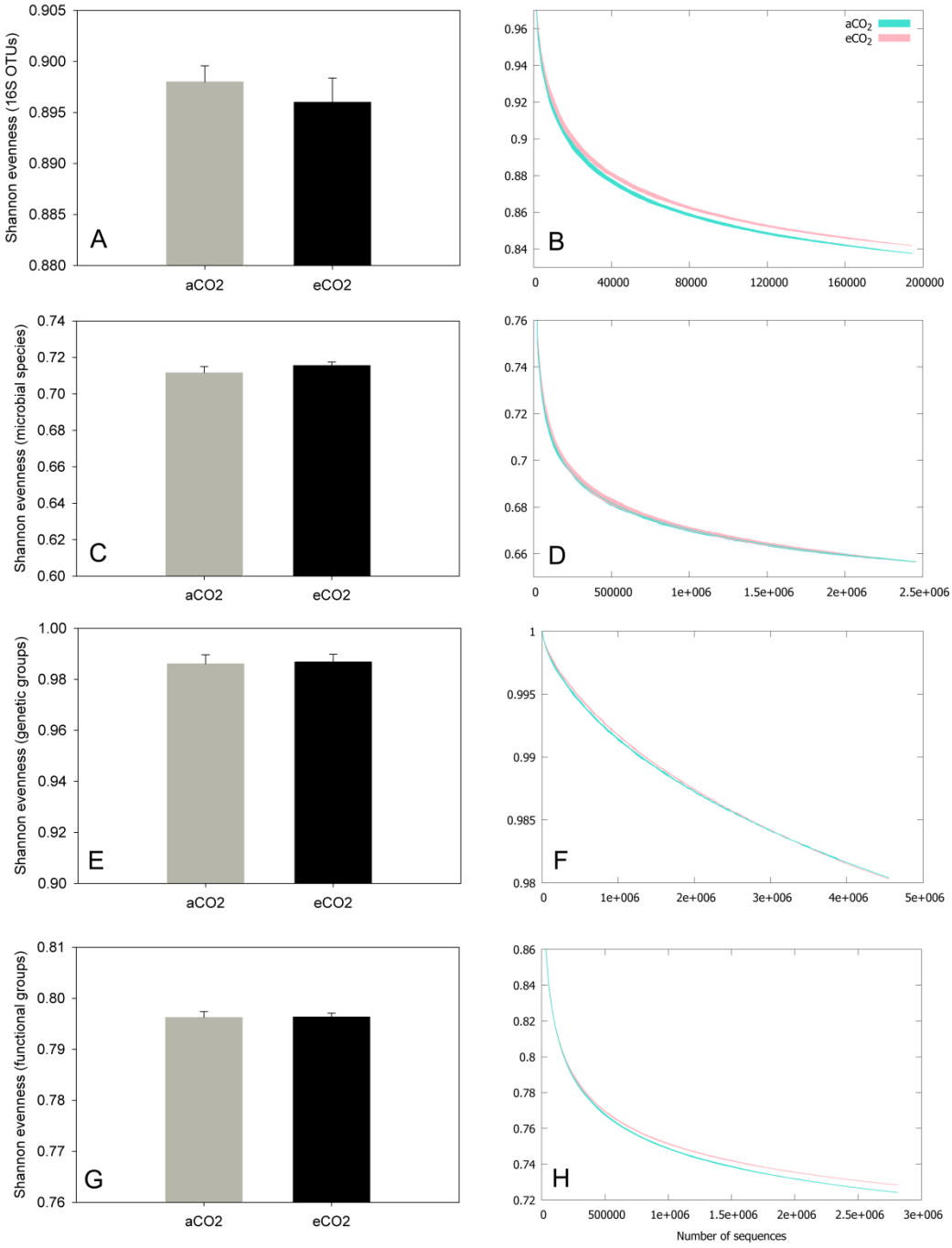


Fig. S3.3 The Shannon evenness of taxonomic (A-D), genetic (E-F) and functional groups in aCO_2 and eCO_2 samples. Both alpha (left) and gamma (right) level evenness was analyzed.

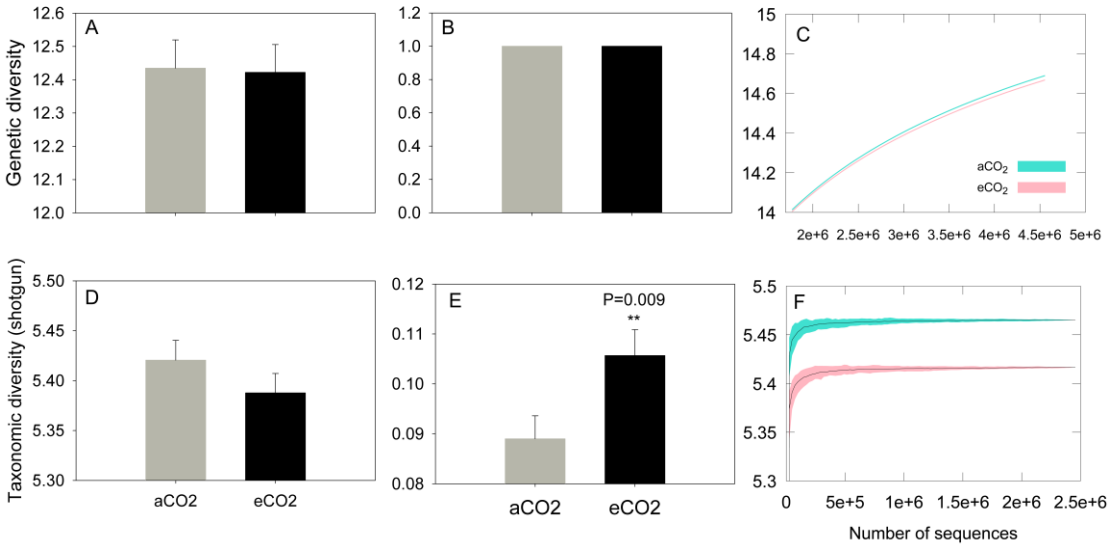


Fig. S3.4 The Shannon diversity for describing genetic and taxonomic diversities identified from shotgun metagenomes. Alpha (left), beta (middle) and gamma (right) diversities were calculated. **: P<0.05.

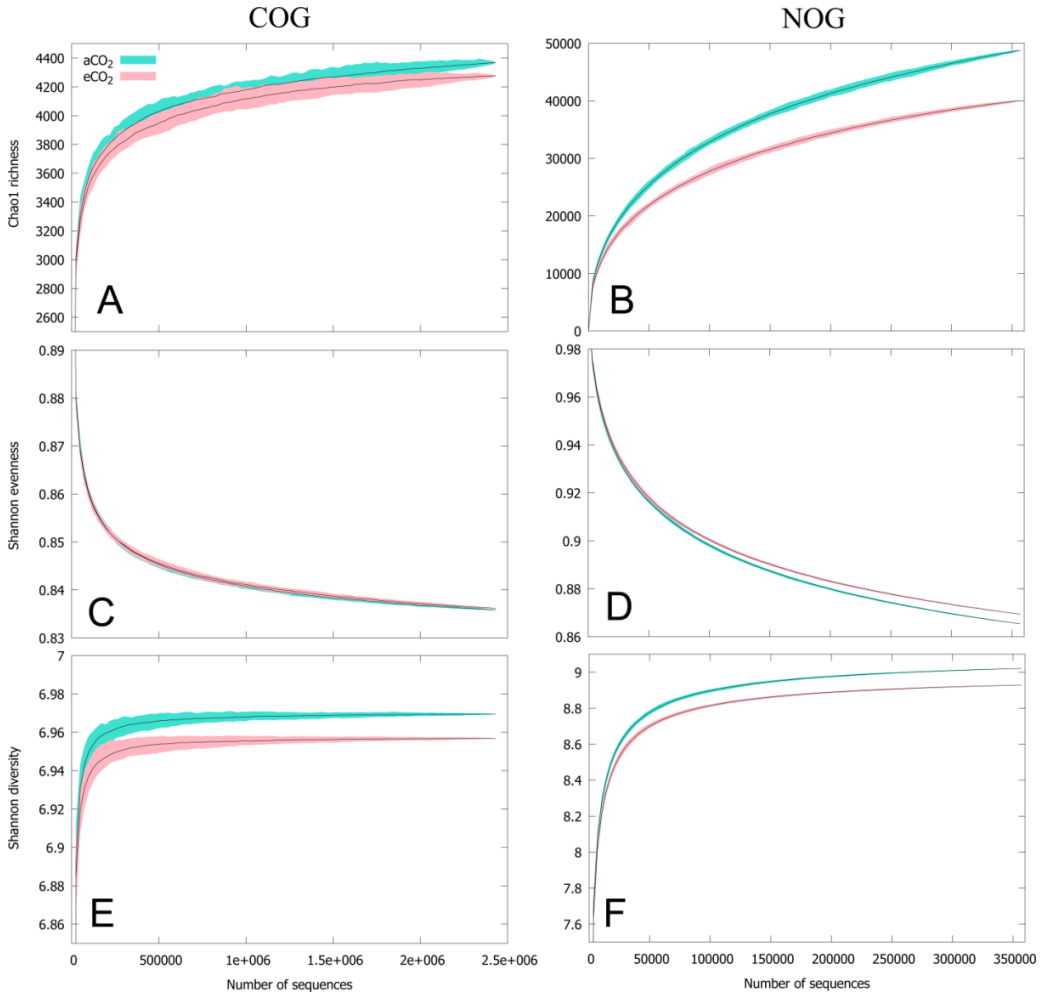


Fig. S3.5 The Chao1 richness (A,B), Shannon evenness (C,D) and Shannon diversity (E,F) for COG (left) and NOG (right) orthologous groups.

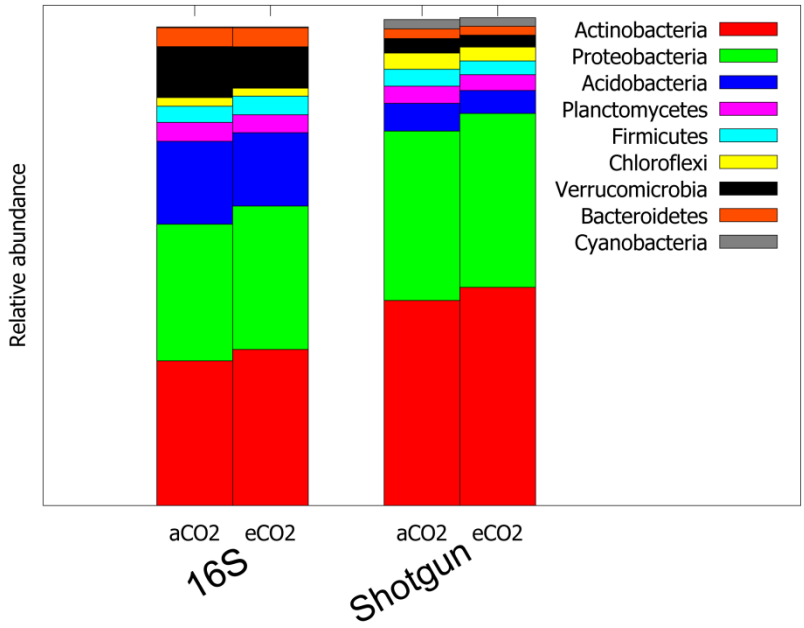


Fig. S3.6 The relative abundance of major bacteria phyla determined by 16S and shotgun metagenomes.

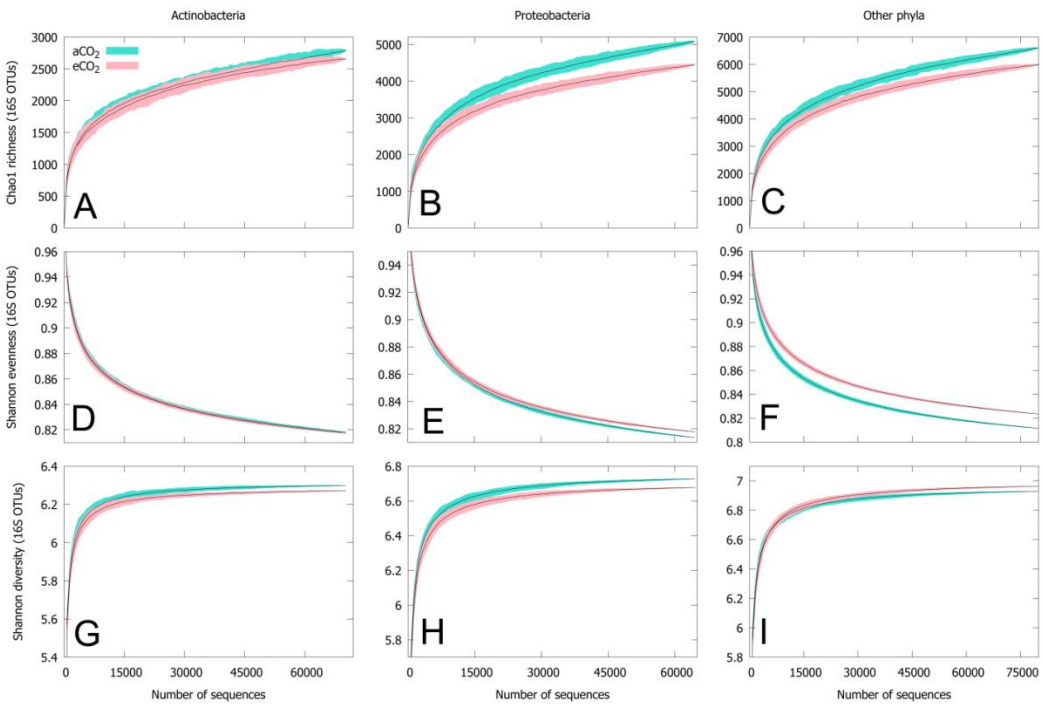


Fig. S3.7 The Chao1 richness (A-C), Shannon evenness (D-F), and Shannon diversity (G-I) for Actinobacteria, Proteobacteria and other rare phyla.

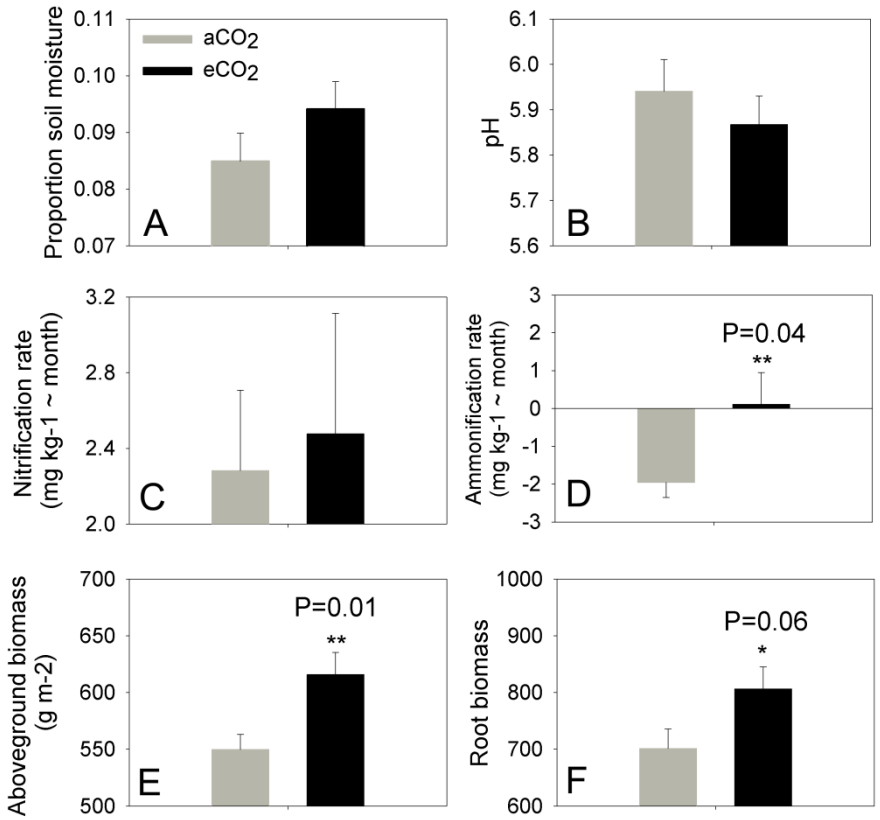


Fig. S3.8 Soil and plant properties under aCO_2 and eCO_2 collected in year 2009. (A) soil moisture, (B) pH, (C) nitrification rate, (D) ammonification, (E) aboveground biomass and (F) root biomass.

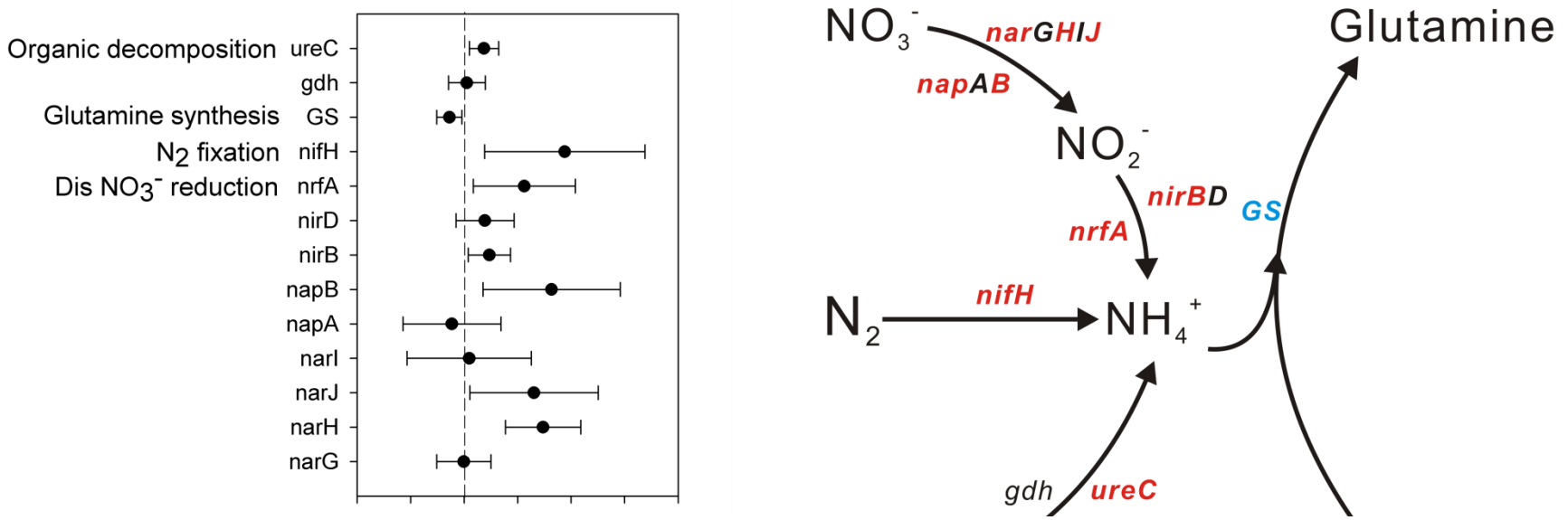


Fig. S3.9 Response ratio analysis of gene families related with NH_4^+ (left). Their relative abundances (middle) and major roles (right) in ammonium pathways are also plotted. Red indicate increased relative abundance, and green indicated decreased relative abundance.

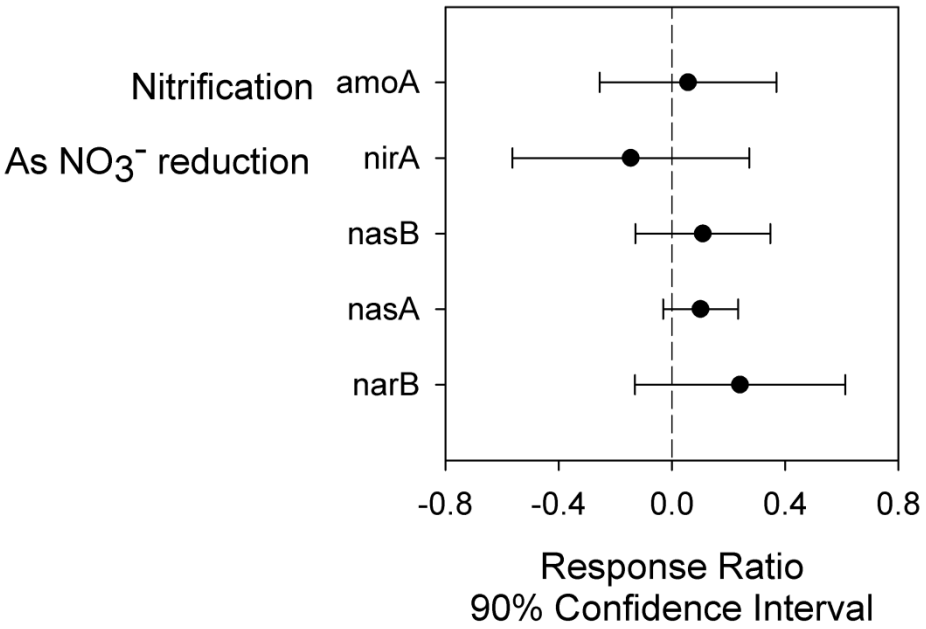


Fig. S3.10 Response ratio analysis of gene families involved in nitrification and assimilatory NO_3^- reduction.

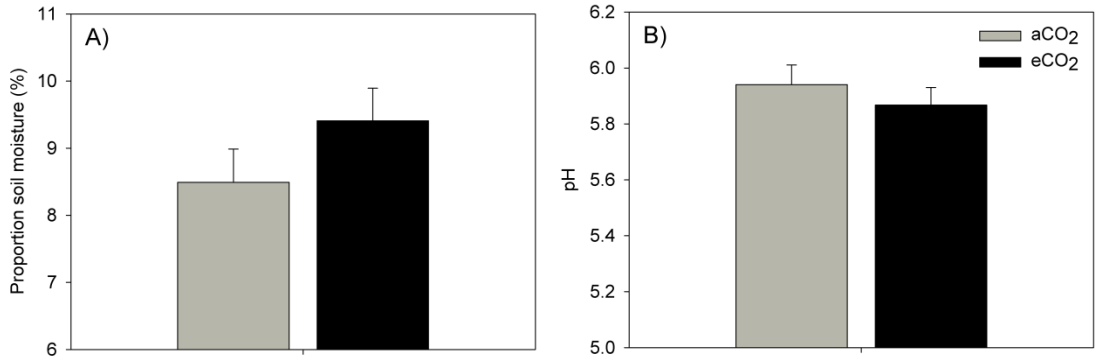


Fig. S4.1 eCO2 effects on soil moisture (A) and soil pH (B). No significant changes were observed for both proportional soil moisture and pH.

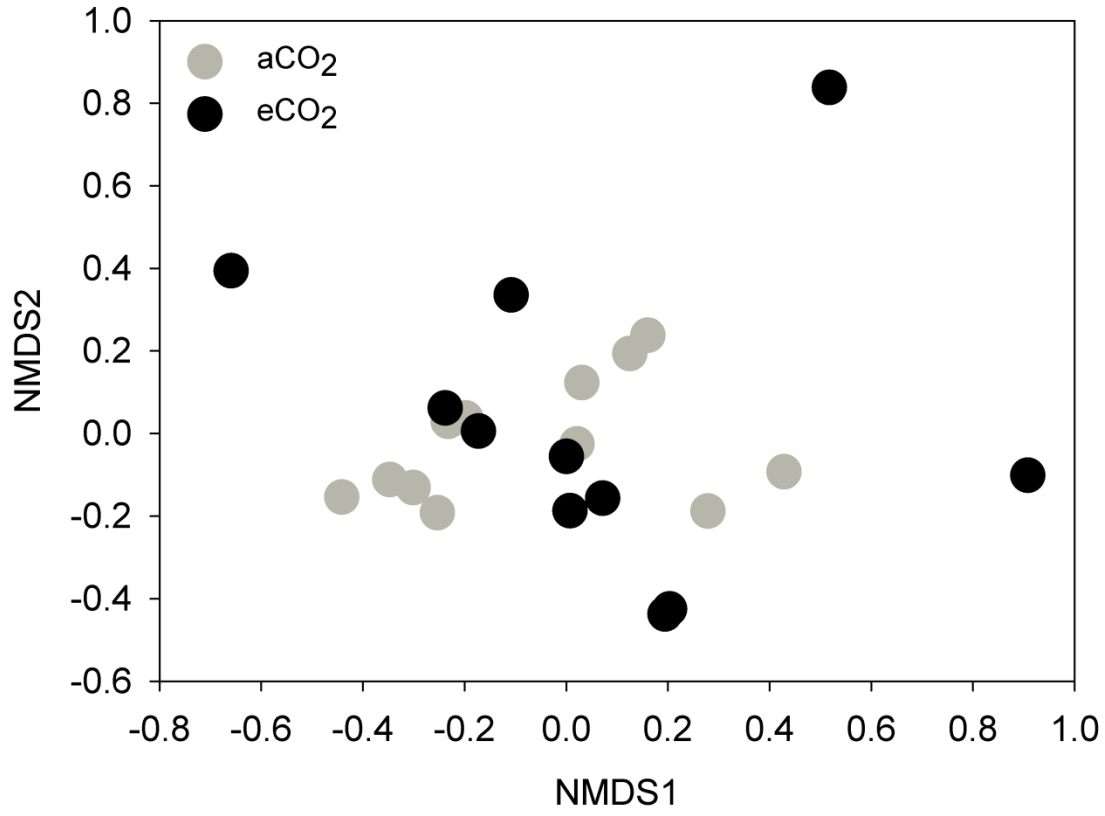


Fig. S4.2 Nonmetric multidimensional scaling analysis of overall fungal community structure under aCO_2 and eCO_2 samples. No clear separations could be observed.

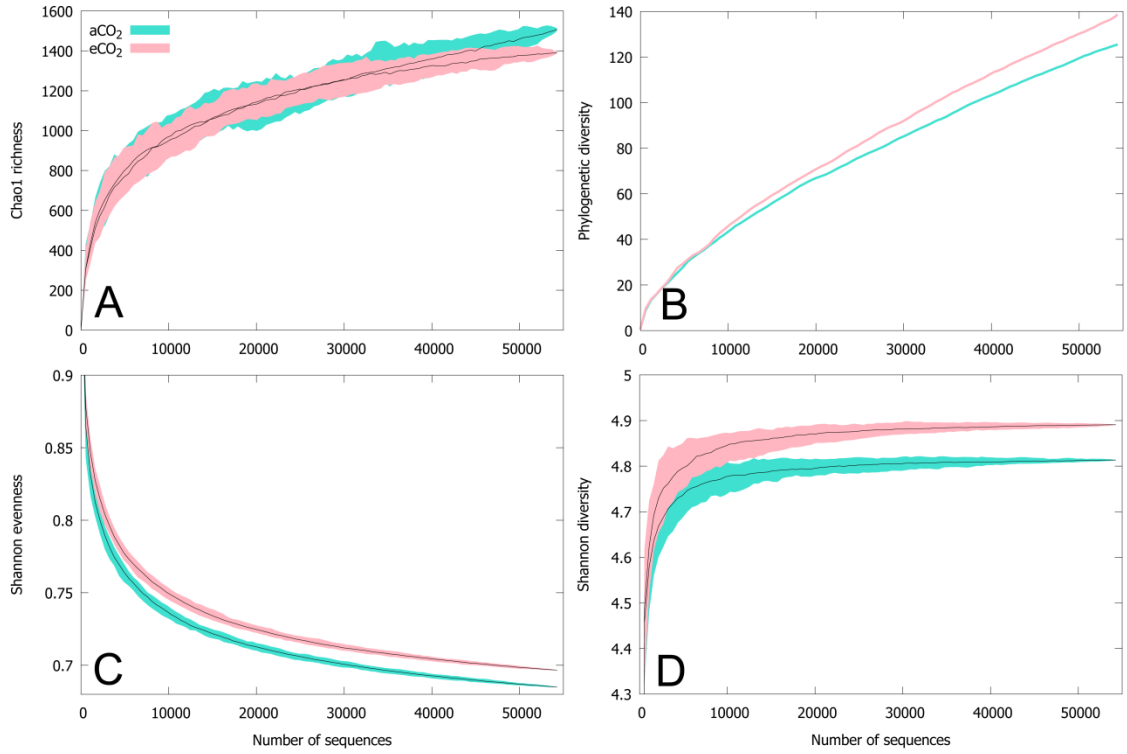


Fig. S4.3 Rarefaction analysis of fungal community species richness (A), phylogenetic diversity (B), species evenness (C), and taxonomic diversity (D) under aCO2 and eCO2 samples. Filled curves refer to 95% confidence intervals.

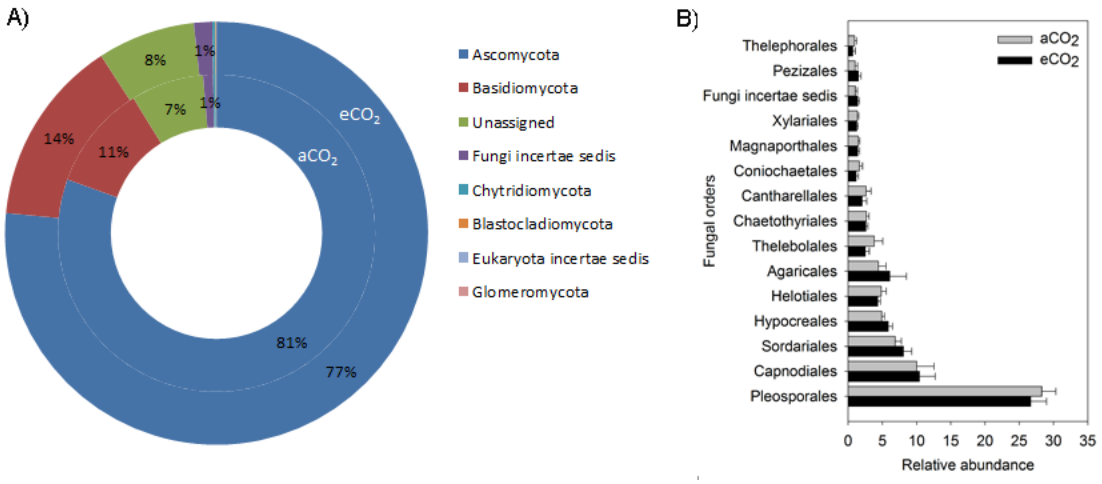


Fig. S4.4 The composition of fungal community at (A) phylum level and (B) order level. Only the top 15 most abundant fungal orders with relative abundance ≥ 0.8% were displayed. These 15 fungal orders accounted for about 75% of the total captured fungal community. Calculation was based on total number of sequences covered by OTUs.

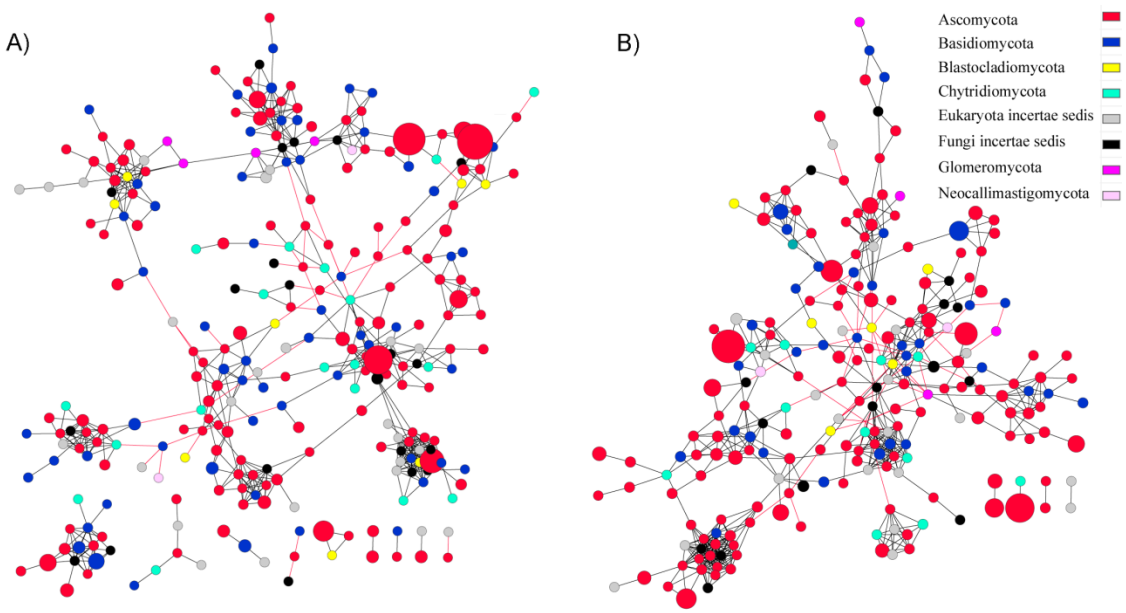


Fig. S4.5 An overview of constructed networks for fungal communities at aCO_2 (A) and eCO_2 (B). More intense connections between nodes and modules were observed in eCO_2 network, showing more complex community interactions under eCO_2. Each node represents an OTU, which could be regarded as a fungal species. The size of nodes represents relative abundance of OTUs. Each link connects two OTUs. Grey links means positive connections, and red means negative connections. Different colors refer to different fungal phyla.

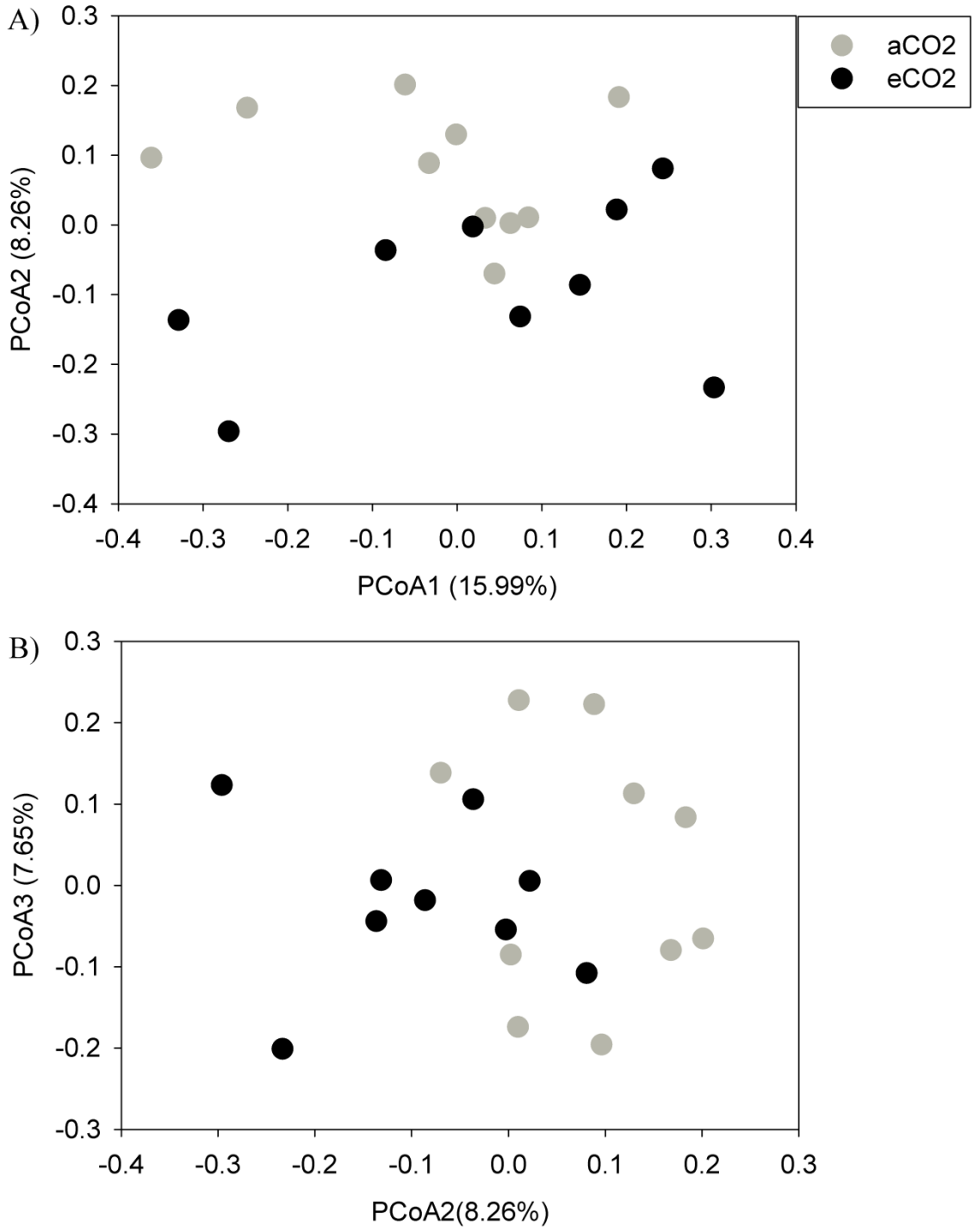


Fig. S5.1 Unweighted UniFrac PCoA analysis of the *nifH* community. A trend of separation was found at all three dimensions analyzed.

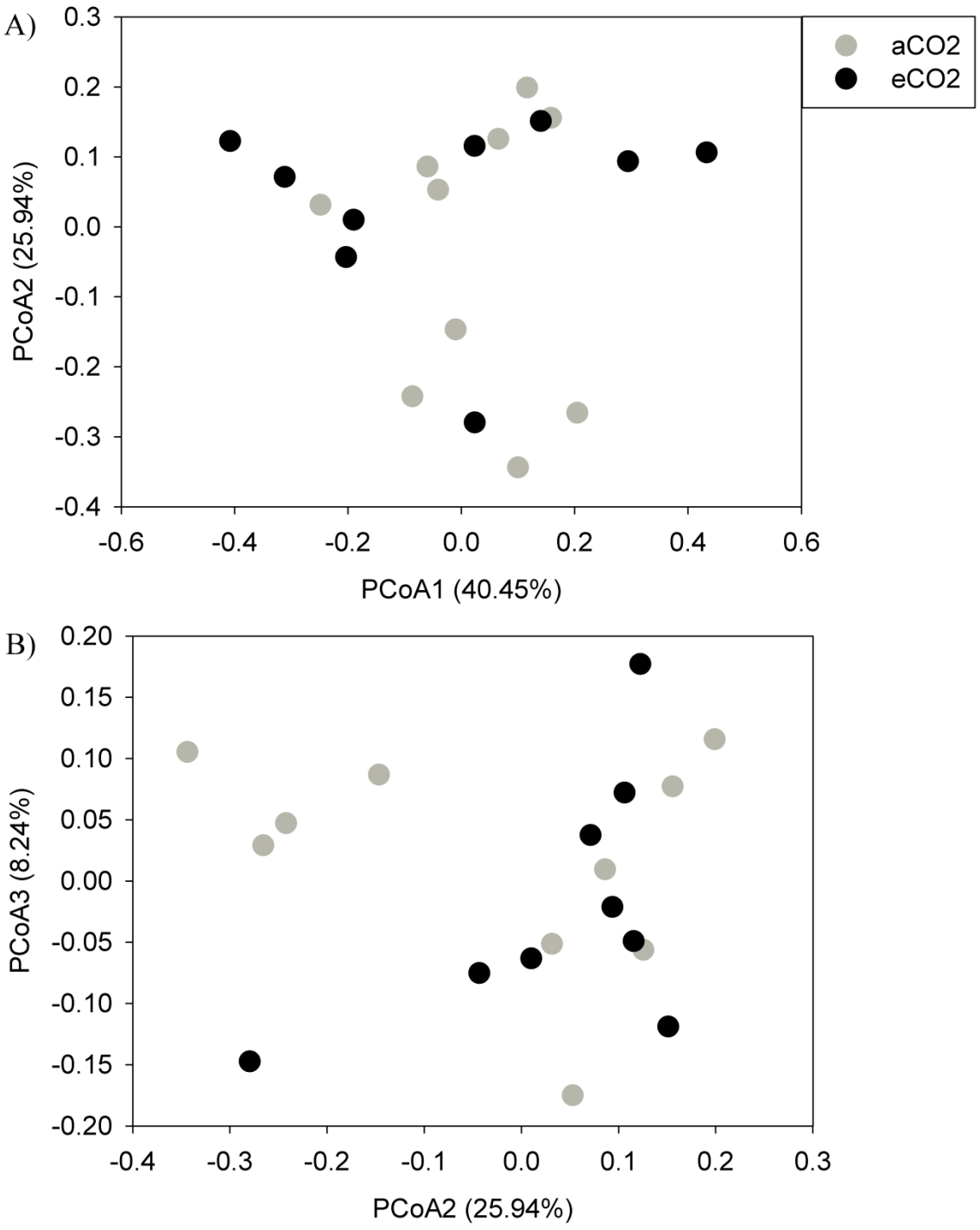


Fig. S5.2 Weighted UniFrac PCoA analysis of the *nifH* community. The trend of separation disappeared when relative abundance of *nifH* OTUs was considered.

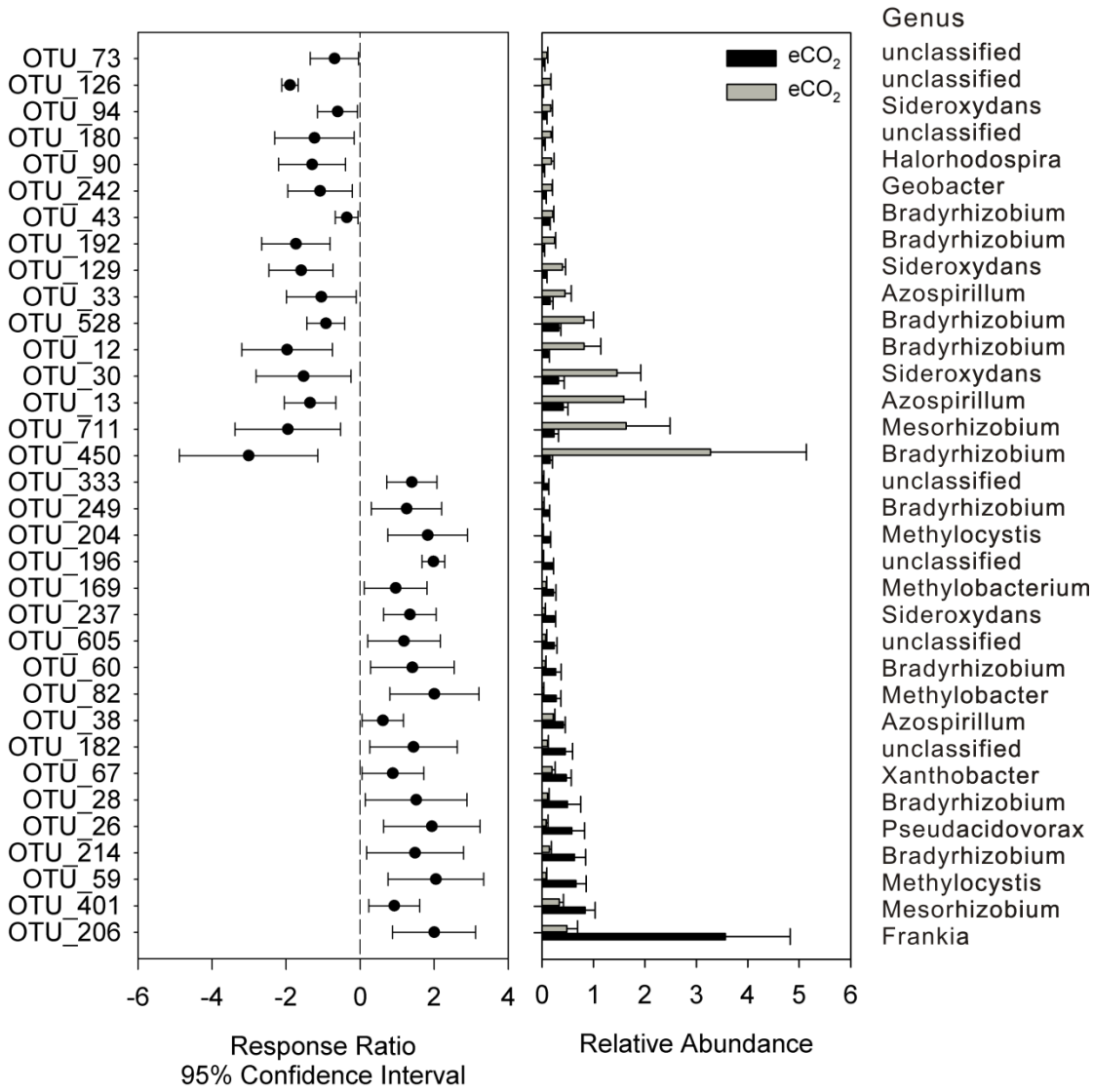


Fig. S5.3 Response ratio analysis of significantly changed *nifH* OTUs. Relative abundance and genus assignment for these OTUs were also included. Error bars plotted at the right side of the dashed line indicate significantly increased relative abundance at eCO_2, while error bars plotted at the left side indicate significantly decreased relative abundance at eCO_2.

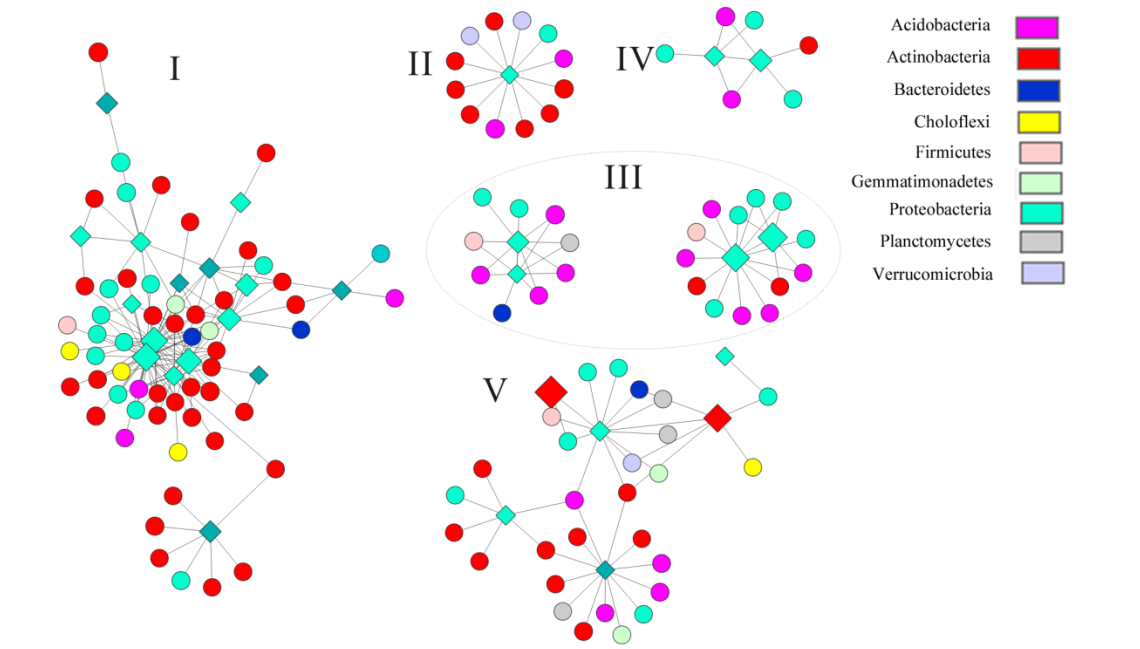


Fig. S5.4 All *nifH*-centered modules identified in this study. Only modules with >5 nodes were included.

References

Ahn, S. J., J. Costa, et al. (1996). "PicoGreen Quantitation of DNA: Effective Evaluation of Samples Pre-or Psost-PCR." Nucleic Acids Research **24**(13): 2623-2625.

Alberton, O., T. W. Kuyper, et al. (2005). "Taking mycocentrism seriously: mycorrhizal fungal and plant responses to elevated CO2." New Phytologist **167**(3): 859-868.

Altschul, S. F., W. Gish, et al. (1990). "Basic local alignment search tool." Journal of molecular biology **215**(3): 403-410.

Angly, F. E., D. Willner, et al. (2012). "Grinder: a versatile amplicon and shotgun sequence simulator." Nucleic Acids Research.

Ansorge, W. J. (2009). "Next-generation DNA sequencing techniques." New Biotechnology **25**(4): 195-203.

Antoninka, A., P. B. Reich, et al. (2011). "Seven years of carbon dioxide enrichment, nitrogen fertilization and plant diversity influence arbuscular mycorrhizal fungi in a grassland ecosystem." New Phytologist **192**(1): 200-214.

Ball, A. (1997). "Microbial decomposition at elevated CO2 levels: effect of litter quality." Global Change Biology **3**(4): 379-386.

Barberan, A., S. T. Bates, et al. (2012). "Using network analysis to explore co-occurrence patterns in soil microbial communities." ISME J **6**(2): 343-351.

Bardgett, R. D., C. Freeman, et al. (2008). "Microbial contributions to climate change through carbon cycle feedbacks." ISME J **2**(8): 805-814.

Barnola, J., D. Raynaud, et al. (1987). Vostok ice core provides 160,000-year record of atmospheric CO2.

Berthrong, S., C. M. Yeager, et al. (2014). "Nitrogen fertilization has a stronger effect on soil N-fixing bacterial communities than elevated atmospheric CO2." Applied and Environmental Microbiology.

Blagodatskaya, E., S. Blagodatsky, et al. (2010). "Elevated atmospheric CO2 increases microbial growth rates in soil: results of three CO2 enrichment experiments." Global Change Biology **16**(2): 836-848.

Bridge, P. and B. Spooner (2001). "Soil fungi: diversity and detection." Plant and Soil **232**(1-2): 147-154.

Brodie, E. L., T. Z. DeSantis, et al. (2006). "Application of a High-Density Oligonucleotide Microarray Approach To Study Bacterial Population Dynamics during Uranium Reduction and Reoxidation." Applied and Environmental Microbiology **72**(9): 6288-6298.

Brodie, E. L., T. Z. DeSantis, et al. (2007). "Urban aerosols harbor diverse and dynamic bacterial populations." Proceedings of the National Academy of Sciences **104**(1): 299-304.

Buée, M., M. Reich, et al. (2009). "454 Pyrosequencing analyses of forest soils reveal an unexpectedly high fungal diversity." New Phytologist **184**(2): 449-456.

Carney, K. M., B. A. Hungate, et al. (2007). "Altered soil microbial community at elevated CO2 leads to loss of soil carbon." Proceedings of the National Academy of Sciences **104**(12): 4990-4995.

Castro, H. F., A. T. Classen, et al. (2010). "Soil Microbial Community Responses to Multiple Experimental Climate Change Drivers." Applied and Environmental Microbiology **76**(4): 999-1007.
Chaffron, S., H. Rehrauer, et al. (2010). "A global network of coexisting microbes from environmental and whole-genome sequence data." Genome Research **20**(7): 947-959.
Chase, J. M., N. J. B. Kraft, et al. (2011). "Using null models to disentangle variation in community dissimilarity from variation in α-diversity." Ecosphere **2**(2): art24.
Cheng, L., F. L. Booker, et al. (2012). "Arbuscular mycorrhizal fungi increase organic carbon decomposition under elevated CO2." Science **337**(6098): 1084-1087.
Chikere, C. B., G. C. Okpokwasili, et al. (2011). "Monitoring of microbial hydrocarbon remediation in the soil." **1**(3): 117-138.
Chou, H. H. and M. H. Holmes (2001). "DNA sequence quality trimming and vector removal." Bioinformatics **17**(12): 1093-1104.
Chung, H., D. Zak, et al. (2006). "Fungal community composition and metabolism under elevated CO2 and O3." Oecologia **147**(1): 143-154.
Chung, H., D. R. Zak, et al. (2007). "Plant species richness, elevated CO2, and atmospheric nitrogen deposition alter soil microbial community composition and function." Global Change Biology **13**(5): 980-989.
Cleveland, C. C., A. R. Townsend, et al. (1999). "Global patterns of terrestrial biological nitrogen (N2) fixation in natural ecosystems." Global Biogeochemical Cycles **13**(2): 623-645.
Collado, M. C., E. Isolauri, et al. (2009). "The impact of probiotic on gut health." Curr Drug Metab **10**(1): 68-78.
Collavino, M. M., H. J. Tripp, et al. (2014). "nifH pyrosequencing reveals the potential for location-specific soil chemistry to influence N2-fixing community dynamics." Environmental Microbiology: n/a-n/a.
Condron, L., C. Stark, et al. (2010). The Role of Microbial Communities in the Formation and Decomposition of Soil Organic Matter. Soil Microbiology and Sustainable Crop Production. G. R. Dixon and E. L. Tilston, Springer Netherlands**:** 81-118.
Conlan, S., H. H. Kong, et al. (2012). "Species-level analysis of DNA sequence data from the NIH Human Microbiome Project." PLoS One **7**(10): 10.
Costello, E. K., C. L. Lauber, et al. (2009). "Bacterial Community Variation in Human Body Habitats Across Space and Time." Science **326**(5960): 1694-1697.
Craine, J. M., C. Morrow, et al. (2007). "Microbial nitrogen limitation increases decomposition." Ecology **88**(8): 2105-2113.
Davidson, E. A. and I. A. Janssens (2006). "Temperature sensitivity of soil carbon decomposition and feedbacks to climate change." Nature **440**(7081): 165-173.
Deng, Y., Z. He, et al. (2012). "Elevated carbon dioxide alters the structure of soil microbial communities." Applied and Environmental Microbiology **78**(8): 2991-2995.
Deng, Y., Z. He, et al. (2012). "Elevated Carbon Dioxide Alters the Structure of Soil Microbial Communities." Applied and Environmental Microbiology **78**(8): 2991-2995.

Deng, Y., Y.-H. Jiang, et al. (2012). "Molecular ecological network analyses." BMC bioinformatics **13**(1): 113.
Drake, J. E., A. Gallet-Budynek, et al. (2011). "Increases in the flux of carbon belowground stimulate nitrogen uptake and sustain the long-term enhancement of forest productivity under elevated CO2." Ecology Letters **14**(4): 349-357.
Drigo, B., G. A. Kowalchuk, et al. (2013). "Impacts of 3 years of elevated atmospheric CO2 on rhizosphere carbon flow and microbial community dynamics." Global Change Biology **19**(2): 621-636.
Drigo, B., G. A. Kowalchuk, et al. (2008). "Climate change goes underground: effects of elevated atmospheric CO2 on microbial community structure and activities in the rhizosphere." Biology and Fertility of Soils **44**(5): 667-679.
Drigo, B., A. S. Pijl, et al. (2010). "Shifting carbon flow from roots into associated microbial communities in response to elevated atmospheric CO2." Proceedings of the National Academy of Sciences **107**(24): 10938-10942.
Drigo, B., J. A. Van Veen, et al. (2009). "Specific rhizosphere bacterial and fungal groups respond differently to elevated atmospheric CO2." The ISME journal **3**(10): 1204-1217.
Dunbar, J., S. A. Eichorst, et al. (2012). "Common bacterial responses in six ecosystems exposed to 10 years of elevated atmospheric carbon dioxide." Environ Microbiol **14**(5): 1145-1158.
Dunbar, J., S. A. Eichorst, et al. (2012). "Common bacterial responses in six ecosystems exposed to 10 years of elevated atmospheric carbon dioxide." Environmental Microbiology **14**(5): 1145-1158.
Eckburg, P. B., E. M. Bik, et al. (2005). "Diversity of the human intestinal microbial flora." Science **308**(5728): 1635-1638.
Eckert, B., O. B. Weber, et al. (2001). "Azospirillum doebereinerae sp. nov., a nitrogen-fixing bacterium associated with the C4-grass Miscanthus." Int J Syst Evol Microbiol **51**(Pt 1): 17-26.
Edgar, R. C. (2004). "MUSCLE: multiple sequence alignment with high accuracy and high throughput." Nucleic Acids Research **32**(5): 1792-1797.
Edgar, R. C. (2013). "UPARSE: highly accurate OTU sequences from microbial amplicon reads." Nature methods.
Edwards, I. P. and D. R. Zak (2011). "Fungal community composition and function after long-term exposure of northern forests to elevated atmospheric CO2 and tropospheric O3." Global Change Biology **17**(6): 2184-2195.
Eisenhauer, N., S. Cesarz, et al. (2012). "Global change belowground: impacts of elevated CO2, nitrogen, and summer drought on soil food webs and biodiversity." Global Change Biology **18**(2): 435-447.
Faith, D. P. (1992). "Conservation evaluation and phylogenetic diversity." Biological Conservation **61**(1): 1-10.
Fath, B. D., U. M. Scharler, et al. (2007). "Ecological network analysis: network construction." Ecological Modelling **208**(1): 49-55.
Faust, K. and J. Raes (2012). "Microbial interactions: from networks to models." Nature Reviews Microbiology **10**(8): 538-550.
Faust, K., J. F. Sathirapongsasuti, et al. (2012). "Microbial Co-occurrence Relationships in the Human Microbiome." PLoS Comput Biol **8**(7): e1002606.

Feng, X., A. J. Simpson, et al. (2010). "Altered microbial community structure and organic matter composition under elevated CO2 and N fertilization in the duke forest." Global Change Biology **16**(7): 2104-2116.
Fierer, N., J. Ladau, et al. (2013). "Reconstructing the Microbial Diversity and Function of Pre-Agricultural Tallgrass Prairie Soils in the United States." Science **342**(6158): 621-624.
Fierer, N., J. W. Leff, et al. (2012). "Cross-biome metagenomic analyses of soil microbial communities and their functional attributes." Proceedings of the National Academy of Sciences.
Fierer, N., M. S. Strickland, et al. (2009). "Global patterns in belowground communities." Ecology Letters **12**(11): 1238-1249.
Finzi, A. C., D. J. Moore, et al. (2006). "Progressive nitrogen limitation of ecosystem processes under elevated CO2 in a warm-temperate forest." Ecology **87**(1): 15-25.
Fischer, H., M. Wahlen, et al. (1999). "Ice core records of atmospheric CO2 around the last three glacial terminations." Science **283**(5408): 1712-1714.
Gaby, J. C. and D. H. Buckley (2011). "A global census of nitrogenase diversity." Environmental Microbiology **13**(7): 1790-1799.
Gaby, J. C. and D. H. Buckley (2014). "A comprehensive aligned nifH gene database: a multipurpose tool for studies of nitrogen-fixing bacteria." Database: the journal of biological databases and curation **2014**.
Galloway, J. N., F. J. Dentener, et al. (2004). "Nitrogen Cycles: Past, Present, and Future." Biogeochemistry **70**(2): 153-226.
Geisseler, D., W. R. Horwath, et al. (2010). "Pathways of nitrogen utilization by soil microorganisms – A review." Soil Biology and Biochemistry **42**(12): 2058-2067.
Gro ßkopf, T., W. Mohr, et al. (2012). "Doubling of marine dinitrogen-fixation rates based on direct measurements." Nature **488**(7411): 361-364.
Groszkopf, T., W. Mohr, et al. (2012). "Doubling of marine dinitrogen-fixation rates based on direct measurements." Nature **488**(7411): 361-364.
Gruber, N. and J. N. Galloway (2008). "An Earth-system perspective of the global nitrogen cycle." Nature **451**(7176): 293-296.
Hamady, M., C. Lozupone, et al. (2009). "Fast UniFrac: facilitating high-throughput phylogenetic analyses of microbial communities including analysis of pyrosequencing and PhyloChip data." The ISME journal **4**(1): 17-27.
Hamer, H. M., D. Jonkers, et al. (2008). "Review article: the role of butyrate on colonic function." Aliment Pharmacol Ther **27**(2): 104-119.
Hatem, A., D. Bozda, et al. (2013). "Benchmarking short sequence mapping tools." BMC Bioinformatics **14**(1): 184.
Hawksworth, D. L. (2001). "The magnitude of fungal diversity: the 1.5 million species estimate revisited." Mycological Research **105**(12): 1422-1432.
Hayden, H. L., P. M. Mele, et al. (2012). "Changes in the microbial community structure of bacteria, archaea and fungi in response to elevated CO2 and warming in an Australian native grassland soil." Environmental Microbiology **14**(12): 3081-3096.

He, Z., Y. Deng, et al. (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., Y. Piceno, et al. (2012). "The phylogenetic composition and structure of soil microbial communities shifts in response to elevated carbon dioxide." ISME J **6**(2): 259-272.
He, Z., M. Xu, et al. (2010). "Metagenomic analysis reveals a marked divergence in the structure of belowground microbial communities at elevated CO2." Ecology Letters **13**(5): 564-575.
He, Z., M. Xu, et al. (2010). "Metagenomic analysis reveals a marked divergence in the structure of belowground microbial communities at elevated CO2." Ecol Lett **13**(5): 564-575.
Heath, J., E. Ayres, et al. (2005). "Rising Atmospheric CO2 Reduces Sequestration of Root-Derived Soil Carbon." Science **309**(5741): 1711-1713.
Hector, A. and R. Bagchi (2007). "Biodiversity and ecosystem multifunctionality." Nature **448**(7150): 188-190.
Hector, A. and R. Hooper (2002). "Darwin and the First Ecological Experiment." Science **295**(5555): 639-640.
Hedges, L. V., J. Gurevitch, et al. (1999). "The meta-analysis of response ratios in experimental ecology." Ecology **80**(4): 1150-1156.
Hemme, C. L., Y. Deng, et al. (2010). "Metagenomic insights into evolution of a heavy metal-contaminated groundwater microbial community." ISME J **4**(5): 660-672.
Hess, M., A. Sczyrba, et al. (2011). "Metagenomic Discovery of Biomass-Degrading Genes and Genomes from Cow Rumen." Science **331**(6016): 463-467.
Hill, M. O. (1973). "Diversity and Evenness: A Unifying Notation and Its Consequences." Ecology **54**(2): 427-432.
Hsu, S.-F. and D. H. Buckley (2009). "Evidence for the functional significance of diazotroph community structure in soil." The ISME journal **3**(1): 124-136.
Hu, S., F. S. Chapin, et al. (2001). "Nitrogen limitation of microbial decomposition in a grassland under elevated CO2." Nature **409**(6817): 188-191.
Human Microbiome Project Consortium (2012). "A framework for human microbiome research." Nature **486**(7402): 215-221.
Human Microbiome Project Consortium (2012). "Structure, function and diversity of the healthy human microbiome." Nature **486**(7402): 207-214.
Huson, D. H., A. F. Auch, et al. (2007). "MEGAN analysis of metagenomic data." Genome research **17**(3): 377-386.
Huson, D. H., S. Mitra, et al. (2011). "Integrative analysis of environmental sequences using MEGAN4." Genome Res **21**(9): 1552-1560.
Izquierdo, J. and K. Nüsslein (2006). "Distribution of Extensive nifH Gene Diversity Across Physical Soil Microenvironments." Microbial Ecology **51**(4): 441-452.
Janda, J. M. and S. L. Abbott (2007). "16S rRNA Gene Sequencing for Bacterial Identification in the Diagnostic Laboratory: Pluses, Perils, and Pitfalls." Journal of Clinical Microbiology **45**(9): 2761-2764.
Janus, L., N. Angeloni, et al. (2005). "Elevated Atmospheric CO2 Alters Soil Microbial Communities Associated with Trembling Aspen (Populus tremuloides) Roots." Microbial Ecology **50**(1): 102-109.

Jasper, J. P. and J. Hayes (1990). "A carbon isotope record of CO2 levels during the late Quaternary." Nature **347**(6292): 462-464.
Jumpponen, A., K. L. Jones, et al. (2010). "Vertical distribution of fungal communities in tallgrass prairie soil." Mycologia **102**(5): 1027-1041.
Kahindi, J., P. Woomer, et al. (1997). "Agricultural intensification, soil biodiversity and ecosystem function in the tropics: the role of nitrogen-fixing bacteria." Applied Soil Ecology **6**(1): 55-76.
Karch, H., P. I. Tarr, et al. (2005). "Enterohaemorrhagic Escherichia coli in human medicine." International Journal of Medical Microbiology **295**(6–7): 405-418.
Kareiva, P. M., J. G. Kingsolver, et al. (1993). Biotic interactions and global change, Sinauer Associates Incorporated.
Karlsson, F. H., V. Tremaroli, et al. (2013). "Gut metagenome in European women with normal, impaired and diabetic glucose control." Nature **498**(7452): 99-103.
Kau, A. L., P. P. Ahern, et al. (2011). "Human nutrition, the gut microbiome and the immune system." Nature **474**(7351): 327-336.
Kent, W. J. (2002). "BLAT—the BLAST-like alignment tool." Genome research **12**(4): 656-664.
Kiehl, J. and K. E. Trenberth (1997). "Earth's annual global mean energy budget." Bulletin of the American Meteorological Society **78**(2): 197-208.
Koike, T., T. Izuta, et al. (1997). Effects of high CO2 on nodule formation in roots of Japanese mountain alder seedlings grown under two nutrient levels. Plant Nutrition for Sustainable Food Production and Environment. T. Ando, K. Fujita, T. Maeet al, Springer Netherlands. **78:** 887-888.
Konstantinidis, K. T. and J. M. Tiedje (2005). "Genomic insights that advance the species definition for prokaryotes." Proceedings of the National Academy of Sciences of the United States of America **102**(7): 2567-2572.
Langley, J. A. and J. P. Megonigal (2010). "Ecosystem response to elevated CO2 levels limited by nitrogen-induced plant species shift." Nature **466**(7302): 96-99.
Larsen, N., F. K. Vogensen, et al. (2010). "Gut Microbiota in Human Adults with Type 2 Diabetes Differs from Non-Diabetic Adults." PLoS One **5**(2): e9085.
Law, C. S., E. Breitbarth, et al. (2012). "No stimulation of nitrogen fixation by non-filamentous diazotrophs under elevated CO2 in the South Pacific." Global Change Biology **18**(10): 3004-3014.
Lesaulnier, C., D. Papamichail, et al. (2008). "Elevated atmospheric CO2 affects soil microbial diversity associated with trembling aspen." Environmental Microbiology **10**(4): 926-941.
Lewin, K. F., G. R. Hendrey, et al. (1994). "Design and application of a free-air carbon dioxide enrichment facility." Agricultural and Forest Meteorology **70**(1): 15-29.
Ley, R. E. (2010). "Obesity and the human microbiome." Current Opinion in Gastroenterology **26**(1): 5-11 10.1097/MOG.1090b1013e328333d328751.
Lin, L., Y. Ji, et al. (2013). "Microevolution from shock to adaptation revealed strategies improving ethanol tolerance and production in Thermoanaerobacter." Biotechnology for biofuels **6**(1): 103.
Lin, L., H. Song, et al. (2011). "The thermoanaerobacter glycobiome reveals mechanisms of pentose and hexose co-utilization in bacteria." PLoS genetics **7**(10): e1002318.

Lipson, D., M. Blair, et al. (2006). "Relationships Between Microbial Community Structure and Soil Processes Under Elevated Atmospheric Carbon Dioxide." Microbial Ecology **51**(3): 302-314.
Lipson, D. A., R. F. Wilson, et al. (2005). "Effects of elevated atmospheric CO2 on soil microbial biomass, activity, and diversity in a chaparral ecosystem." Applied and Environmental Microbiology **71**(12): 8573-8580.
Liu, K.-L., A. Porras-Alfaro, et al. (2012). "Accurate, Rapid Taxonomic Classification of Fungal Large-Subunit rRNA Genes." Applied and Environmental Microbiology **78**(5): 1523-1533.
Loreau, M., S. Naeem, et al. (2001). "Biodiversity and Ecosystem Functioning: Current Knowledge and Future Challenges." Science **294**(5543): 804-808.
Lozupone, C. and R. Knight (2005). "UniFrac: a new phylogenetic method for comparing microbial communities." Applied and Environmental Microbiology **71**(12): 8228-8235.
Lozupone, C. A., M. Hamady, et al. (2008). "The convergence of carbohydrate active gene repertoires in human gut microbes." Proceedings of the National Academy of Sciences **105**(39): 15076-15081.
Luo, F., Y. Yang, et al. (2007). "Constructing gene co-expression networks and predicting functions of unknown genes by random matrix theory." BMC bioinformatics **8**(1): 299.
Luo, Y., D. Hui, et al. (2006). "ELEVATED CO2 STIMULATES NET ACCUMULATIONS OF CARBON AND NITROGEN IN LAND ECOSYSTEMS: A META-ANALYSIS." Ecology **87**(1): 53-63.
Luo, Y., B. O. Su, et al. (2004). "Progressive Nitrogen Limitation of Ecosystem Responses to Rising Atmospheric Carbon Dioxide." BioScience **54**(8): 731-739.
Lyons, K. G., C. A. Brigham, et al. (2005). "Rare Species and Ecosystem Functioning." Conservation Biology **19**(4): 1019-1024.
Müller, T., B. Walter, et al. (2006). "Ammonium Toxicity in Bacteria." Current Microbiology **52**(5): 400-406.
Mackelprang, R., M. P. Waldrop, et al. (2011). "Metagenomic analysis of a permafrost microbial community reveals a rapid response to thaw." Nature **480**(7377): 368-371.
MacLean, D., J. D. G. Jones, et al. (2009). "Application of 'next-generation' sequencing technologies to microbial genetics." Nat Rev Micro **7**(4): 287-296.
Magoč, T. and S. L. Salzberg (2011). "FLASH: Fast Length Adjustment of Short Reads to Improve Genome Assemblies." Bioinformatics.
Marçais, G. and C. Kingsford (2011). "A fast, lock-free approach for efficient parallel counting of occurrences of k-mers." Bioinformatics **27**(6): 764-770.
Mardis, E. R. (2008). "The impact of next-generation sequencing technology on genetics." Trends in genetics **24**(3): 133.
Martinez-Romero, E. (2009). "Coevolution in Rhizobium-legume symbiosis?" DNA Cell Biol **28**(8): 361-370.
McCann, K. S. (2000). "The diversity-stability debate." Nature **405**(6783): 228-233.
McDonald, D., M. N. Price, et al. (2012). "An improved Greengenes taxonomy with explicit ranks for ecological and evolutionary analyses of bacteria and archaea." ISME J **6**(3): 610-618.

Meinzer, F. C. (2003). "Functional convergence in plant responses to the environment." Oecologia **134**(1): 1-11.
Metzker, M. L. (2010). "Sequencing technologies [mdash] the next generation." Nat Rev Genet **11**(1): 31-46.
Mohamed, N. M., A. S. Colman, et al. (2008). "Diversity and expression of nitrogen fixation genes in bacterial symbionts of marine sponges." Environmental Microbiology **10**(11): 2910-2921.
Moisander, P. H., L. Shiue, et al. (2006). "Application of a nifH oligonucleotide microarray for profiling diversity of N2-fixing microorganisms in marine microbial mats." Environmental Microbiology **8**(10): 1721-1735.
Mokany, K., J. Ash, et al. (2008). "Functional identity is more important than diversity in influencing ecosystem processes in a temperate native grassland." Journal of Ecology **96**(5): 884-893.
Muller, J., D. Szklarczyk, et al. (2010). "eggNOG v2. 0: extending the evolutionary genealogy of genes with enhanced non-supervised orthologous groups, species and functional annotations." Nucleic Acids Research **38**(suppl 1): D190-D195.
Muller, J., D. Szklarczyk, et al. (2010). "eggNOG v2.0: extending the evolutionary genealogy of genes with enhanced non-supervised orthologous groups, species and functional annotations." Nucleic Acids Res **38**(Database issue): 9.
Musso, G., R. Gambino, et al. (2011). "Interactions between gut microbiota and host metabolism predisposing to obesity and diabetes." Annu Rev Med **62**: 361-380.
Mutch, L. A. and J. P. Young (2004). "Diversity and specificity of Rhizobium leguminosarum biovar viciae on wild and cultivated legumes." Mol Ecol **13**(8): 2435-2444.
Naeem, S., J. E. Duffy, et al. (2012). "The Functions of Biological Diversity in an Age of Extinction." Science **336**(6087): 1401-1406.
Norby, R. J., E. H. DeLucia, et al. (2005). "Forest response to elevated CO2 is conserved across a broad range of productivity." Proceedings of the National Academy of Sciences of the United States of America **102**(50): 18052-18056.
Norby, R. J. and Y. Luo (2004). "Evaluating ecosystem responses to rising atmospheric CO2 and global warming in a multi-factor world." New Phytologist **162**(2): 281-293.
Norby, R. J., J. M. Warren, et al. (2010). "CO2 enhancement of forest productivity constrained by limited nitrogen availability." Proceedings of the National Academy of Sciences **107**(45): 19368-19373.
O'Brien, H. E., J. L. Parrent, et al. (2005). "Fungal Community Analysis by Large-Scale Sequencing of Environmental Samples." Applied and Environmental Microbiology **71**(9): 5544-5550.
Okubo, T., T. Tsukui, et al. (2012). "Complete genome sequence of Bradyrhizobium sp. S23321: insights into symbiosis evolution in soil oligotrophs." Microbes Environ **27**(3): 306-315.
ÖPik, M., M. Moora, et al. (2006). "Composition of root-colonizing arbuscular mycorrhizal fungal communities in different ecosystems around the globe." Journal of Ecology **94**(4): 778-790.
Parker, M. A. (2012). "Legumes select symbiosis island sequence variants in Bradyrhizobium." Molecular Ecology **21**(7): 1769-1778.

Parrent, J. L., W. F. Morris, et al. (2006). "CO2-ENRICHMENT AND NUTRIENT AVAILABILITY ALTER ECTOMYCORRHIZAL FUNGAL COMMUNITIES." Ecology **87**(9): 2278-2287.
Parrent, J. L. and R. Vilgalys (2007). "Biomass and compositional responses of ectomycorrhizal fungal hyphae to elevated CO2 and nitrogen fertilization." New Phytologist **176**(1): 164-174.
Paruelo, J. M., E. G. Jobbágy, et al. (1998). "Functional and structural convergence of temperate grassland and shrubland ecosystems." Ecological Applications **8**(1): 194-206.
Penton, C. R., D. St. Louis, et al. (2013). "Fungal Diversity In Permafrost And Tallgrass Prairie Soils Under Experimental Warming." Applied and Environmental Microbiology.
Peterson, J., S. Garges, et al. (2009). "The NIH Human Microbiome Project." Genome Res **19**(12): 2317-2323.
Phillips, R. P., I. C. Meier, et al. (2012). "Roots and fungi accelerate carbon and nitrogen cycling in forests exposed to elevated CO2." Ecology Letters **15**(9): 1042-1049.
Poly, F., L. J. Monrozier, et al. (2001). "Improvement in the RFLP procedure for studying the diversity of< i> nifH</i> genes in communities of nitrogen fixers in soil." Research in Microbiology **152**(1): 95-103.
Price, M. N., P. S. Dehal, et al. (2009). "FastTree: computing large minimum evolution trees with profiles instead of a distance matrix." Molecular biology and evolution **26**(7): 1641-1650.
Pruesse, E., C. Quast, et al. (2007). "SILVA: a comprehensive online resource for quality checked and aligned ribosomal RNA sequence data compatible with ARB." Nucleic Acids Research **35**(21): 7188-7196.
Pryde, S. E., S. H. Duncan, et al. (2002). "The microbiology of butyrate formation in the human colon." FEMS Microbiol Lett **217**(2): 133-139.
Qin, J., R. Li, et al. (2010). "A human gut microbial gene catalogue established by metagenomic sequencing." Nature **464**(7285): 59-65.
Qin, J., Y. Li, et al. (2012). "A metagenome-wide association study of gut microbiota in type 2 diabetes." Nature **490**(7418): 55-60.
Rappe, M. S. and S. J. Giovannoni (2003). "The uncultured microbial majority." Annu Rev Microbiol **57**: 369-394.
Ravel, J., P. Gajer, et al. (2010). "Vaginal microbiome of reproductive-age women." Proceedings of the National Academy of Sciences.
Raymond, J., J. L. Siefert, et al. (2004). "The Natural History of Nitrogen Fixation." Molecular biology and evolution **21**(3): 541-554.
Raynaud, D. and J. Barnola (1985). "An Antarctic ice core reveals atmospheric CO2 variations over the past few centuries." Nature **315**(6017): 309-311.
Reich, P. B. and S. E. Hobbie (2013). "Decade-long soil nitrogen constraint on the CO2 fertilization of plant biomass." Nature Clim. Change **3**(3): 278-282.
Reich, P. B., S. E. Hobbie, et al. (2006). "Nitrogen limitation constrains sustainability of ecosystem response to CO2." Nature **440**(7086): 922-925.
Reich, P. B., S. E. Hobbie, et al. (2006). "Nitrogen limitation constrains sustainability of ecosystem response to CO2." Nature **440**(7086): 922-925.

Reich, P. B., J. Knops, et al. (2001). "Plant diversity enhances ecosystem responses to elevated CO2 and nitrogen deposition." Nature **410**(6830): 809-810.
Reich, P. B., J. Knops, et al. (2001). "Plant diversity enhances ecosystem responses to elevated CO2 and nitrogen deposition." Nature **410**(6830): 809-810.
Reich, P. B., D. Tilman, et al. (2012). "Impacts of Biodiversity Loss Escalate Through Time as Redundancy Fades." Science **336**(6081): 589-592.
Reich, P. B., M. B. Walters, et al. (1997). "From tropics to tundra: global convergence in plant functioning." Proceedings of the National Academy of Sciences **94**(25): 13730-13734.
Reinhold, B., T. Hurek, et al. (1987). "Azospirillum halopraeferens sp. nov., a Nitrogen-Fixing Organism Associated with Roots of Kallar Grass (Leptochloa fusca (L.) Kunth)." International Journal of Systematic Bacteriology **37**(1): 43-51.
Rho, M., H. Tang, et al. (2010). "FragGeneScan: predicting genes in short and error-prone reads." Nucleic Acids Res **38**(20): 30.
Robinson, T., G. McMullan, et al. (2001). "Remediation of dyes in textile effluent: a critical review on current treatment technologies with a proposed alternative." Bioresource Technology **77**(3): 247-255.
Salyers, A. A., A. Gupta, et al. (2004). "Human intestinal bacteria as reservoirs for antibiotic resistance genes." Trends in microbiology **12**(9): 412-416.
Santos-Gonz áez, J. C., R. D. Finlay, et al. (2007). "Seasonal Dynamics of Arbuscular Mycorrhizal Fungal Communities in Roots in a Seminatural Grassland." Applied and Environmental Microbiology **73**(17): 5613-5623.
Schatz, M. C., A. M. Phillippy, et al. (2010). "Integrated microbial survey analysis of prokaryotic communities for the PhyloChip microarray." Applied and Environmental Microbiology **76**(16): 5636-5638.
Schell, M. A., M. Karmirantzou, et al. (2002). "The genome sequence of Bifidobacterium longum reflects its adaptation to the human gastrointestinal tract." Proc Natl Acad Sci U S A **99**(22): 14422-14427.
Schloss, P. D., S. L. Westcott, et al. (2009). "Introducing mothur: open-source, platform-independent, community-supported software for describing and comparing microbial communities." Applied and Environmental Microbiology **75**(23): 7537-7541.
Scholz, M. B., C.-C. Lo, et al. (2012). "Next generation sequencing and bioinformatic bottlenecks: the current state of metagenomic data analysis." Current opinion in biotechnology **23**(1): 9-15.
Scholz, M. B., C.-C. Lo, et al. (2012). "Next generation sequencing and bioinformatic bottlenecks: the current state of metagenomic data analysis." Current Opinion in Biotechnology **23**(1): 9-15.
Schwabe, R. F. and C. Jobin (2013). "The microbiome and cancer." Nat Rev Cancer **13**(11): 800-812.
Scott, L. J., K. L. Mohlke, et al. (2007). "A genome-wide association study of type 2 diabetes in Finns detects multiple susceptibility variants." Science **316**(5829): 1341-1345.
Segata, N., S. Haake, et al. (2012). "Composition of the adult digestive tract bacterial microbiome based on seven mouth surfaces, tonsils, throat and stool samples." Genome Biology **13**(6): R42.

Segata, N., L. Waldron, et al. (2012). "Metagenomic microbial community profiling using unique clade-specific marker genes." Nat Meth **9**(8): 811-814.
Sela, D. A., J. Chapman, et al. (2008). "The genome sequence of Bifidobacterium longum subsp. infantis reveals adaptations for milk utilization within the infant microbiome." Proceedings of the National Academy of Sciences **105**(48): 18964-18969.
Shabalina, S. A., A. Y. Ogurtsov, et al. (2001). "Selective constraint in intergenic regions of human and mouse genomes." Trends in Genetics **17**(7): 373-376.
Shaver, G., L. Street, et al. (2007). "Functional convergence in regulation of net CO2 flux in heterogeneous tundra landscapes in Alaska and Sweden." Journal of Ecology **95**(4): 802-817.
Shendure, J. and H. Ji (2008). "Next-generation DNA sequencing." Nat Biotech **26**(10): 1135-1145.
Smith, M. D. and A. K. Knapp (2003). "Dominant species maintain ecosystem function with non-random species loss." Ecology Letters **6**(6): 509-517.
Smoot, M. E., K. Ono, et al. (2011). "Cytoscape 2.8: new features for data integration and network visualization." Bioinformatics **27**(3): 431-432.
Solomon, S. (2007). Climate change 2007-the physical science basis: Working group I contribution to the fourth assessment report of the IPCC, Cambridge University Press.
Song, Y., C. Liu, et al. (2004). "Real-time PCR quantitation of clostridia in feces of autistic children." Applied and Environmental Microbiology **70**(11): 6459-6465.
Sridhar, J., R. Sabarinathan, et al. (2011). "Junker: An Intergenic Explorer for Bacterial Genomes." Genomics, Proteomics & Bioinformatics **9**(4–5): 179-182.
Stacey, G. (1995). "Bradyrhizobium japonicum nodulation genetics." FEMS Microbiology Letters **127**(1-2): 1-9.
Stanish, L. F., S. P. O'Neill, et al. (2013). "Bacteria and diatom co-occurrence patterns in microbial mats from polar desert streams." Environ Microbiol **15**(4): 1115-1131.
Steele, J. A., P. D. Countway, et al. (2011). "Marine bacterial, archaeal and protistan association networks reveal ecological linkages." ISME J **5**(9): 1414-1425.
Steenhoudt, O. and J. Vanderleyden (2000). "Azospirillum, a free-living nitrogen-fixing bacterium closely associated with grasses: genetic, biochemical and ecological aspects." FEMS Microbiol Rev **24**(4): 487-506.
Stocker, D. Q. (2013). "Climate change 2013: The physical science basis." Working Group I Contribution to the Fifth Assessment Report of the Intergovernmental Panel on Climate Change, Summary for Policymakers, IPCC.
Suding, K. N., S. L. Collins, et al. (2005). "Functional- and abundance-based mechanisms explain diversity loss due to N fertilization." Proceedings of the National Academy of Sciences of the United States of America **102**(12): 4387-4392.
Tien, T. M., M. H. Gaskins, et al. (1979). "Plant Growth Substances Produced by Azospirillum brasilense and Their Effect on the Growth of Pearl Millet (Pennisetum americanum L.)." Applied and Environmental Microbiology **37**(5): 1016-1024.

Tilman, D., P. B. Reich, et al. (2006). "Biodiversity and ecosystem stability in a decade-long grassland experiment." Nature **441**(7093): 629-632.
Toju, H., A. S. Tanabe, et al. (2012). "High-Coverage ITS Primers for the DNA-Based Identification of Ascomycetes and Basidiomycetes in Environmental Samples." PLoS ONE **7**(7): e40863.
Tringe, S. G., C. von Mering, et al. (2005). "Comparative Metagenomics of Microbial Communities." Science **308**(5721): 554-557.
Tu, Q., H. Yu, et al. (2014). "GeoChip 4: a functional gene-array-based high-throughput environmental technology for microbial community analysis." Molecular ecology resources.
Turnbaugh, P. J., M. Hamady, et al. (2009). "A core gut microbiome in obese and lean twins." Nature **457**(7228): 480-484.
Turnbaugh, P. J., M. Hamady, et al. (2009). "A core gut microbiome in obese and lean twins." Nature **457**(7228): 480-484.
Tylianakis, J. M., R. K. Didham, et al. (2008). "Global change and species interactions in terrestrial ecosystems." Ecology Letters **11**(12): 1351-1363.
van der Heijden, M. G., J. N. Klironomos, et al. (1998). "Mycorrhizal fungal diversity determines plant biodiversity, ecosystem variability and productivity." Nature **396**(6706): 69-72.
Vane-Wright, R. I., C. J. Humphries, et al. (1991). "What to protect?—Systematics and the agony of choice." Biological Conservation **55**(3): 235-254.
Venter, J. C., K. Remington, et al. (2004). "Environmental Genome Shotgun Sequencing of the Sargasso Sea." Science **304**(5667): 66-74.
Verbruggen, E., S. D. Veresoglou, et al. (2013). "Arbuscular mycorrhizal fungi–short-term liability but long-term benefits for soil carbon storage?" New Phytologist **197**(2): 366-368.
Victoria Wang, X., N. Blades, et al. (2012). "Estimation of sequencing error rates in short reads." BMC Bioinformatics **13**(1): 185.
Wagg, C., S. F. Bender, et al. (2014). "Soil biodiversity and soil community composition determine ecosystem multifunctionality." Proceedings of the National Academy of Sciences.
Walther, G.-R., E. Post, et al. (2002). "Ecological responses to recent climate change." Nature **416**(6879): 389-395.
Wang, Q., G. M. Garrity, et al. (2007). "Naive Bayesian classifier for rapid assignment of rRNA sequences into the new bacterial taxonomy." Applied and Environmental Microbiology **73**(16): 5261-5267.
Wang, Q., J. F. Quensen, et al. (2013). "Ecological Patterns of nifH Genes in Four Terrestrial Climatic Zones Explored with Targeted Metagenomics Using FrameBot, a New Informatics Tool." mBio **4**(5).
Wang, X. F. and G. Chen (2003). "Complex networks: small-world, scale-free and beyond." Circuits and Systems Magazine, IEEE **3**(1): 6-20.
Wang, Y. and P.-Y. Qian (2009). "Conservative Fragments in Bacterial 16S rRNA Genes and Primer Design for 16S Ribosomal DNA Amplicons in Metagenomic Studies." PLoS One **4**(10): e7401.
Wardle, D. A., R. D. Bardgett, et al. (2004). "Ecological linkages between aboveground and belowground biota." Science **304**(5677): 1629-1633.

Weber, C. F., R. Vilgalys, et al. (2013). "Changes in fungal community composition in response to elevated atmospheric CO2 and nitrogen fertilization varies with soil horizon." Frontiers in Microbiology **4**.
Webster, J. and R. Weber (2007). Introduction to Fungi, Cambridge University Press.
Wellen, K. E. and G. S. Hotamisligil (2005). "Inflammation, stress, and diabetes." J Clin Invest **115**(5): 1111-1119.
Wexler, H. M. (2007). "Bacteroides: the Good, the Bad, and the Nitty-Gritty." Clinical Microbiology Reviews **20**(4): 593-621.
Wu, G. D., J. Chen, et al. (2011). "Linking Long-Term Dietary Patterns with Gut Microbial Enterotypes." Science **334**(6052): 105-108.
Xu, L., S. Ravnskov, et al. (2012). "Linking fungal communities in roots, rhizosphere, and soil to the health status of Pisum sativum." FEMS Microbiology Ecology **82**(3): 736-745.
Xu, M., Z. He, et al. (2013). "Elevated CO2 influences microbial carbon and nitrogen cycling." BMC microbiology **13**(1): 124.
Yang, Y., D. P. Harris, et al. (2008). "Characterization of the Shewanella oneidensis Fur gene: roles in iron and acid tolerance response." BMC genomics **9**(Suppl 1): S11.
Young, J. (1992). "Phylogenetic classification of nitrogen-fixing organisms." Biological nitrogen fixation: 43-86.
Zak, D. R., K. S. Pregitzer, et al. (2011). "Forest productivity under elevated CO2 and O3: positive feedbacks to soil N cycling sustain decade-long net primary productivity enhancement by CO2." Ecology Letters **14**(12): 1220-1226.
Zavaleta, E. S., J. R. Pasari, et al. (2010). "Sustaining multiple ecosystem functions in grassland communities requires higher biodiversity." Proceedings of the National Academy of Sciences.
Zehr, J. P. (2011). "Nitrogen fixation by marine cyanobacteria." Trends in microbiology **19**(4): 162-173.
Zehr, J. P., B. D. Jenkins, et al. (2003). "Nitrogenase gene diversity and microbial community structure: a cross-system comparison." Environ Microbiol **5**(7): 539-554.
Zehr, J. P., B. D. Jenkins, et al. (2003). "Nitrogenase gene diversity and microbial community structure: a cross-system comparison." Environmental Microbiology **5**(7): 539-554.
Zehr, J. P., M. Mellon, et al. (1995). "Diversity of heterotrophic nitrogen fixation genes in a marine cyanobacterial mat." Applied and Environmental Microbiology **61**(7): 2527-2532.
Zhang, Z., S. Schwartz, et al. (2000). "A greedy algorithm for aligning DNA sequences." Journal of Computational Biology **7**(1-2): 203-214.
Zhou, A., Z. He, et al. (2010). "Hydrogen peroxide-induced oxidative stress responses in Desulfovibrio vulgaris Hildenborough." Environmental Microbiology **12**(10): 2645-2657.
Zhou, J., M. A. Bruns, et al. (1996). "DNA recovery from soils of diverse composition." Applied and Environmental Microbiology **62**(2): 316-322.
Zhou, J., Y. Deng, et al. (2010). "Functional Molecular Ecological Networks." mBio **1**(4).

Zhou, J., Y. Deng, et al. (2011). "Phylogenetic Molecular Ecological Network of Soil Microbial Communities in Response to Elevated CO2." mBio **2**(4).
Zhou, J., K. Xue, et al. (2012). "Microbial mediation of carbon-cycle feedbacks to climate warming." Nature Clim. Change **2**(2): 106-110.