Title

The effects of anthropogenic disturbance and environmental change on multiple dimensions of microbial biodiversity

Author

Doll, Hannah Mariah

Publication Date

2016

Peer reviewed|Thesis/dissertation

The effects of anthropogenic disturbance and environmental change on multiple dimensions of microbial biodiversity

by

Hannah Mariah Doll

A thesis submitted in partial satisfaction of the

requirements for the degree of

Doctor of Philosophy

in

Environmental Science, Policy, and Management

in the

Graduate Division

of the

University of California, Berkeley

Committee in charge:

Professor Matthew D. Potts, Chair

Professor Mary K. Firestone

Professor Brent D. Mishler

Professor Hélène Morlon

Summer 2016

The effects of anthropogenic disturbance and environmental change on multiple dimensions of microbial biodiversity

by Hannah Mariah Doll

ABSTRACT

The effects of anthropogenic disturbance and environmental

change on multiple dimensions of microbial biodiversity

by

Hannah Mariah Doll

Doctor of Philosophy in Environmental Science, Policy, and Management

University of California, Berkeley

Professor Matthew D. Potts, Chair

Despite recent advances in microbial ecology, including the widespread use of high-throughput sequencing and micro-array technologies, microbial taxonomic, phylogenetic, and functional diversity remain understudied and poorly understood compared to our knowledge of macrofaunal diversity. In this dissertation, I work to close this gap by: (1) addressing the challenges of quantifying and comparing modern microbial data, (2) elucidating the changes to soil microbial diversity caused by widespread land-use change in tropical ecosystems, and (3) modeling the unique evolution of marine diatoms.

The dissertation begins in Chapter 1 with a general overview of the three original research projects that were carried out. In Chapter 2, I explore the use of diversity profiles, which are a novel way to analyze microbial datasets. Diversity profiles may be better suited than traditional ecological indices for quantifying data spanning multiple domains of life and dimensions of diversity. I evaluate the use of diversity profiles for analyzing microbial assemblages in order to determine whether the inclusion of rarity and similarity information changes the interpretation of comparative studies of microbial community diversity.

In Chapter 3, I assess the effects of anthropogenic land-use change, soil abiotic factors, and geographic distance on the taxonomic, phylogenetic, and functional gene diversity of soil microbes. I discover and quantify multiple dimensions of bacterial, archaeal, and fungal diversity in five different land-use types (Primary Forest, Secondary Forest, Oil Palm, Rubber, and Rice) throughout a dipterocarp forest landscape in Peninsular Malaysia. In Chapter 4, I identify major shifts in lineage diversification rates during diatom evolution by building a new diatom phylogenetic tree with significantly more environmental diatom sequences than previously published phylogenies.

The dissertation concludes in Chapter 5 with a summary of key findings: Microbial diversity comparisons may vary when taxa rarity and similarity information are considered by diversity profiles. Incorporating this information can greatly alter our comparisons and

conclusions of microbial diversity in multi-community studies (Chapter 2); conversion of Primary Forest to other land-use types led to the loss of rare microbial OTUs (Chapter 3); fungal diversity was more strongly affected by land-use type than bacterial and archaeal diversity (Chapter 3); functional gene diversity was most strongly linked to abiotic soil environment (Chapter 3); and analyses of the global diatom phylogenetic tree yield estimates of diversification rate shifts across the tree with all but one of the estimated shifts corresponding to net increases in diversification rates (Chapter 4).

To my family

TABLE OF CONTENTS

ACKNOWLEDGEMENTS

First and foremost, I would like to thank my advisor, Matthew Potts, for his tremendous support during the past seven years. He has been a steadfast source of advice, feedback, constructive criticism, and encouragement, and he has been willing to sit down to talk about my work every single time I have asked him. His dedication to his students and his true hope that we succeed is unparalleled. I am incredibly grateful for the experience of learning from him as a scientist, as well as for the opportunity to study at Berkeley.

I would like to thank my dissertation committee members, Hélène Morlon, Mary Firestone, and Brent Mishler for their guidance, mentorship, feedback, and patience. I would especially like to thank Hélène for sponsoring me in France as a Chateaubriand Fellow and for encouraging and supporting me in the application process and in all of my endeavors while in France. I would also like to thank Paul Fine and Kevin O'Hara for serving on my qualifying exam committee and for helping me refine my initial research ideas.

I would like to thank my labmates in the Potts Group, especially Siew Chin Chua, Matthew Luskin, Benjamin Ramage, Lisa Kelly, Jenny Palomino, Tara DiRocco, Teny Issakhanian, Samuel Evans, Jessica Bryant, and Megan Bartlett for their camaraderie, expertise, and helpful feedback. The memories I have made working with them and the things I have learned from them, particularly during our time in Malaysia, have impacted me greatly.

I would like to thank my collaborators and co-authors for each of my dissertation chapters, who were deeply involved in many aspects of project design, data collection and analysis, and writing. Chapter 2 was co-authored with David Armitage, Rebecca Daly, Joanne Emerson, Daniela Aliaga Goltsman, Alexis Yelton, Jennifer Kerekes, Mary Firestone, and Matthew Potts. For Chapter 3, I collaborated with Jizhong Zhou, Tzeng Yih Lam, Mary Firestone, and Matthew Potts. For Chapter 4, I worked closely with Hélène Morlon, Matthew Potts, Chris Bowler, Colomban de Vargas, Lucie Bittner, and Shruti Malviya. I would also like to thank the entire Tara Oceans research team for allowing me to access the incredibly exciting Tara Oceans diatom dataset. Without these generous collaborators, none of this dissertation would have come to fruition.

I would like to thank the numerous people and institutions that have provided project planning advice, practical support, and help in the field during the course of this dissertation. These include Christine Fletcher, Jaya Radha Veerasamy, Lee Su See, Sadali Sahat, Abd Rahman Kassim, Shamsudin Ibrahim, Forest Research Institute Malaysia, and the Pasoh Research Council for their assistance in obtaining research permits for soil sampling in Peninsular Malaysia, for facilitating site visits, and for their input into the study design of Chapter 3. I would also like to thank all of the landowners and land managers for allowing us to sample soil on their land, specifically Ms. Leong and Ms. Lim at Cap Asas Pls, the Lam family, and especially Chang Kwai Lam for his expertise and time spent guiding us through his rubber and oil palm plantations. I would like to thank Katerina Estera, Shengjing Shi, Rebecca Daly, Joy Van Nostrand, and Sydney Glassman for their helpful input into soil extraction and purification methods and data analysis. Lastly, I would like to thank Sandy Andelman, Julia Parrish, Christina Maranto, Rachel Sewell Nesteruk, James Prosser, Tom Bruns, and all of the other Dimensions of Biodiversity Distributed Graduate Seminar participants for their constructive advice about Chapter 2.

Funding for my salary and for my dissertation research was provided by a National Science Foundation Graduate Research Fellowship, a University of California, Berkeley Chancellor's Fellowship, a Chateaubriand Fellowship granted by the French Embassy to the United States, a University of California, Berkeley Committee on Research grant, and a National Science Foundation Grant (#1050680) to Sandy Andelman and Julia Parrish: The Dimensions of Biodiversity Distributed Graduate Seminar. I am extremely grateful to these institutions for the opportunity to pursue such exciting and interesting research questions for my dissertation.

I would like to thank my family, Mary, Sabri, Frank, Rachel, Drew, Cleo, Nicholas, Marie, and Mark for their love, never-ending support, and patience while I have worked on my doctoral degree for the past seven years. I would like to thank my dad for giving me his love for the outdoors, Rachel for her example of being a great scientist and mom, Drew for his graduate school and academia related advice, and Frank for encouraging me to take a Bioinformatics class at Pomona and for all of his help with Bioinformatics-related computer challenges over the years. I would especially like to thank my mom for her dedication to helping me finish this dissertation and for the countless hours she spent entertaining Mark with boundless enthusiasm.

Last, but certainly not least, I would like to thank my husband, Mike, and my son, Mark. Mike has been incredibly patient and encouraging, while I have been in graduate school working on my dissertation the entire time we have known each other. I would not have been able to finish this dissertation without his unwavering support and love. While I may have finished this dissertation earlier and more easily without Mark, he has brought me immeasurable joy these past two years.

CHAPTER 1. Dissertation Overview

Ecologists have studied plant and animal diversity and the roles these macro-organisms play in the environment for centuries (Haeckel 1866, Hagen 1992). Through these studies, we have garnered valuable knowledge regarding the roles different organisms play in critical ecosystem services, such as maintaining healthy ecosystems. Numerous studies also demonstrate that humans benefit greatly from protecting nature (Balmford et al. 2002, De Groot et al. 2002, Postel and Thompson 2005, Worm et al. 2006, Nelson et al. 2009, Isbell et al. 2011).

Microbes have also long attracted attention from ecologists (Winogradsky 1887, Beijerinck 1888, Baas-Becking 1934). It is well known that microbes play many crucial roles in emergent ecosystem processes (Torsvik and Øvreås 2002), including decomposition (Setälä 2004), nutrient cycling (Arrigo 2005), metal remediation (Valls and De Lorenzo 2002), driving plant diversity and productivity (Van Der Heikden et a. 2008), and the mediation of cycles of the most important atmospherically reactive trace gases (Watling and Harper 1998). However, until recently, microbial ecological research was limited to studying species that could be directly observed and morphologically identified in the environment (i.e., some fungi) or measured with older laboratory methods (e.g., clone libraries, low-depth sequencing). These methodological limitations meant that ecologists had been unable to study an incredibly large swath of overall total global biodiversity.

With the advent of high-throughout sequencing and micro-array technologies, microbial ecologists are now able to uncover much more information about microbial communities than was previously accessible. Illumina HiSeq 3000/HiSeq 4000 systems can now generate up to 1.5 Tb and 5 billion reads per run (Illumina 2016). These metagenomic data have allowed ecologists to address for the first time numerous questions regarding microbial community taxonomic, phylogenetic, and functional diversity (Xu 2006, Mardis 2011, Willner and Hugenholtz 2013, Pershina et al. 2013, Stephens et al. 2015). New insights into microbial community diversity have already been incredibly valuable, having already served to inform marine and land-use management and conservation (Azam and Worden 2004, Besemer et al. 2013, Thomsen and Willerslev 2015), food production (Alkema et al. 2016, Bokulich et al. 2016), sewage treatment (Rizzo et al. 2013), biotechnology (Rashid and Stingl 2015), biomedical research (Dominguez-Bello et al. 2016), and air quality management in hospitals (Kembel et al. 2012).

Current challenges in microbial ecology

Despite recent advances in microbial ecology, including the widespread use of high-throughput sequencing and micro-array technologies, microbial taxonomic, phylogenetic, and functional diversity still remain understudied and poorly understood compared to our knowledge of macrofaunal diversity. New high-throughput technologies do not erase the fact that microbial ecological research still faces several challenges due to microbes' small size, particular evolutionary history, and reproductive mechanisms, which are all very different than those of macro-organisms (Taylor et al. 1999, Eisen 2000, Hughes et al. 2001, Bowler et al. 2008, Chan et al. 2012, Jousset et al. 2013).

In this dissertation, my first focus is addressing the challenges of quantifying and comparing modern microbial community datasets. Microbial ecologists often employ long-

standing methods from classical community ecology to analyze microbial community diversity. While the use of established ecological metrics to analyze microbial diversity may sometimes be appropriate (Hill et al. 2003), the data produced by ecologists surveying macro-organismal communities differ from data obtained by high-throughput sequencing of microbial communities in three key ways. First, in contrast to plant and animal assemblages, microbial assemblages are typically made up of more than one domain of life, thus necessitating the ability to quantify diversity across very disparate organism types. Second, many classical indices assume ecological communities are composed of unique species. However, traditional biological species concepts do not fit the natural histories of many microbial taxa that routinely undergo non-homologous recombination and sometimes lack sexual reproduction (Taylor et al. 2000, Rosselló-Mora and Amann 2001, Staley 2006). The concept of species is widely questioned for macro-organisms as well (Mishler 2010). Finally, unlike with macro-organisms, researchers are often unable to directly observe and characterize microbes and their traits in situ (Tiedje et al. 1999, Luo et al. 2007). Instead, the taxonomic/phylogenetic and functional genes of environmental microbes are commonly sequenced. However, it remains difficult to link the taxonomy of an individual microbe to the environmental functions it carries out.

The second point of focus of this dissertation is elucidating the changes to soil microbial diversity caused by widespread land-use change in tropical ecosystems. Southeast Asia has a very high rate of deforestation with the total forest cover of the region dropping from 268 Mha in 1990 to 236 Mha in 2010 (Stibig et al. 2014). By 2100, the region may have lost three quarters of its primary forests (Sodhi et al. 2004). In addition to having major impacts on plant and animal diversity, tropical land-use change may be directly catalyzing changes to emergent ecosystem processes mediated by soil microbes. Soil bacteria, archaea, and fungi provide a number of crucial ecosystem functions in forests, including driving nutrient cycling, decomposition, soil food webs, plant diversity and productivity, and the cycles of the most important atmospherically reactive trace gases that contribute to an increase in the radiative forcing of our atmosphere (Watling and Harper 1998, Van Der Heijden et al. 2008, de Vries et al. 2013, Cleveland et al. 2014). Despite the crucial roles soil microbes play in forest ecosystems, many questions remain unanswered regarding how multiple dimensions of microbial diversity (i.e., taxonomic, phylogenetic, functional diversity) react to major anthropogenic land-use changes (e.g., logging, conversion to agriculture) in tropical regions.

The third focus of this dissertation is modeling the unique evolution of marine diatoms. Diatoms are the most diverse group of marine phytoplankton, and they carry out several crucial ecosystem services, including about one fifth of total global photosynthesis. Due to their extreme diversity and globally important ecosystem functions, the evolutionary history of diatom diversification is of key interest. Exploring why speciation and extinction rates of organisms vary over time, space, and different types of organisms is crucial to understanding how biological diversity is generated (Morlon 2014). The evolutionary history of diatoms has been studied using the diatom fossil record (Rabosky and Sorhannus 2009, Cermeno et al. 2015, Finkel et al. 2005, Lazarus et al. 2014). However, even though the diatom fossil record is rich, it suffers from sampling biases. Besides fossils, the evolutionary history of diversification can also be studied using phylogenies inferred from molecular data. Diversification trajectories and correlates of speciation and extinction rates can be inferred from molecular phylogenies by utilizing recently developed modeling (reviewed in Morlon 2014). While some recent studies have inferred diatom phylogenies from molecular data (Theriot et al. 2009, Theriot et al. 2010), these studies have focused on deep resolutions, rather than on inferring a species level diatom phylogeny. In

addition to the need to sample additional diatoms in the environment in order to infer a more accurate diatom phylogeny, the dynamics of diatom diversification have not yet been explored using phylogenetic comparative methods.

Overview of Chapters

In Chapter 2, I evaluate the use of diversity profiles for analyzing microbial assemblages in order to determine whether the inclusion of rarity information and similarity data (phylogenetic data, in particular) changes the interpretation of experimental and observational microbial data. Diversity profiles (Leinster and Cobbold 2012) are a novel, promising way to analyze microbial datasets that may be better suited for data spanning multiple domains of life and dimensions of diversity than traditional ecological indices. Diversity profiles encompass many other indices, provide effective numbers of diversity (mathematical generalizations of previous indices that better convey the magnitude of differences in diversity), and can incorporate taxa similarity information. To explore whether these profiles change interpretations of microbial datasets, diversity profiles were calculated for four microbial datasets from different environments spanning all domains of life as well as viruses. Both similarity-based profiles that incorporated phylogenetic relatedness and naïve (not similarity-based) profiles were calculated. Simulated datasets were also used to examine the sensitivity of diversity profiles to varying phylogenetic topology and community composition.

In Chapter 3, I discover and quantify multiple dimensions of bacterial, archaeal, and fungal taxonomic and functional diversity in five different land-use types (Primary Forest, Secondary Forest, Oil Palm, Rubber, and Rice) throughout a dipterocarp forest landscape in Peninsular Malaysia. The specific objectives of this work were to: (1) Assess the effects of anthropogenic land-use change on bacterial, archaeal, and fungal taxonomic diversity in soils from these five different land-use types; (2) Investigate relationships between soil microbial taxonomic diversity and local environment and spatial distance; and (3) Investigate how land-use type, soil abiotic factors, and geographic distance affect the functional gene diversity of soil microbes. Soil samples were collected in the five different land-use types throughout a dipterocarp forest landscape in Peninsular Malaysia. For each of the land-use types, soil samples were taken at three different sampling sites separated by at least 1 km. 16S and ITS Illumina MiSeq-based sequencing and GeoChip 5.0_60K functional gene array (containing approximately 160,000 probes from 378,000 genes) analyses were then performed on the samples. For each sampling site, soil texture (percentage sand, silt, and clay), total nitrogen content, total carbon content, extractable phosphorus (Bray method), and cation exchange capacity were also determined. The resulting data were analyzed to quantify the effects of land-use type, soil abiotic characteristics, and geographic distance on multiple dimensions of soil bacterial, archaeal, and fungal diversity. Data analyses included the use of diversity profiles, cluster dendrograms, detrended correspondence analyses, canonical correspondence analyses, mantel correlations, and Faith's phylogenetic diversity calculations.

In Chapter 4, I investigate diatom evolutionary history by building a new diatom phylogenetic tree with significantly more environmental diatom sequences than in previously published phylogenies. I then use this new phylogeny to identify major shifts in lineage diversification rates (significant increases or decreases in speciation and extinction rates) during the evolution of diatoms. In order to carry out this study, I utilized the Tara Oceans dataset. The

Tara Oceans expedition was an unprecedented effort that led to the sampling of 35,000 samples from 210 ocean sampling sites that contained millions of marine organisms study. Within the Tara Oceans sampling, microscopic plankton were sampled methodically at 210 sites at depths of up to 2000m in all major ocean regions from 2009 to 2013, resulting in 63,371 diatom barcoding sequences (Bork et al. 2015). I combined these Tara Oceans diatom sequences with published diatom databases to build a phylogenetic tree of global diatom diversity and then investigate diatom diversification dynamics using Modeling Evolutionary Diversification Using Stepwise AIC (MEDUSA).

In Chapter 5, I summarize the key findings of this dissertation and discuss potential future research directions. Overall, the dissertation explores the effects of anthropogenic disturbance and environmental change on multiple dimensions of microbial biodiversity. It provides insight into how microbial diversity evolves, how microbial diversity reacts to environmental change, and the methods that allow us to quantitatively study and compare the diversity of different microbial communities. Thus, the dissertation demonstrates how the world's rapidly changing environment affects microbial diversity and serves as a jumping-off point for additional studies investigating the effects of global change on microbial communities. This is especially relevant given the critical roles that microbial diversity plays in provisioning ecosystem services, and the fact that recent advances in technology now allow us to investigate microbial communities in novel ways.

CHAPTER 2. Utilizing novel diversity estimators to quantify multiple dimensions of microbial biodiversity across domains

Abstract

Microbial ecologists often employ methods from classical community ecology to analyze microbial community diversity. However, these methods have limitations because microbial communities differ from macro-organismal communities in key ways. This study sought to quantify microbial diversity using methods that are better suited for data spanning multiple domains of life and dimensions of diversity. Diversity profiles are one novel, promising way to analyze microbial datasets. Diversity profiles encompass many other indices, provide effective numbers of diversity (mathematical generalizations of previous indices that better convey the magnitude of differences in diversity), and can incorporate taxa similarity information. To explore whether these profiles change interpretations of microbial datasets, diversity profiles were calculated for four microbial datasets from different environments spanning all domains of life as well as viruses. Both similarity-based profiles that incorporated phylogenetic relatedness and naïve (not similarity-based) profiles were calculated. Simulated datasets were used to examine the robustness of diversity profiles to varying phylogenetic topology and community composition.

Diversity profiles provided insights into microbial datasets that were not detectable with classical univariate diversity metrics. For all datasets analyzed, there were key distinctions between calculations that incorporated phylogenetic diversity as a measure of taxa similarity and naïve calculations. The profiles also provided information about the effects of rare species on diversity calculations. Additionally, diversity profiles were used to examine thousands of simulated microbial communities, showing that similarity-based and naïve diversity profiles only agreed approximately 50% of the time in their classification of which sample was most diverse. This is a strong argument for incorporating similarity information and calculating diversity with a range of emphases on rare and abundant species when quantifying microbial community diversity.

For many datasets, diversity profiles provided a different view of microbial community diversity compared to analyses that did not take into account taxa similarity information, effective diversity, or multiple diversity metrics. These findings are a valuable contribution to data analysis methodology in microbial ecology.

Introduction

With the widespread use of culture-independent, high-throughput sequencing technologies, ecologists have begun to describe the diversity of microbial communities that were previously difficult to detect (e.g., Roesch et al. 2007, Fulthorpe et al. 2008, Fierer et al. 2011). Given the newness of these data types and the fact that the aims and goals of microbial studies are usually similar to those of macro-ecology, microbial ecologists often use methods from classical community ecology to analyze their data. These include Shannon's H (Shannon 1948), Berger-Parker Evenness (Berger and Parker 1970), rarefaction, and ordination (Bent and Forney 2008).

While the use of established ecological metrics to analyze microbial diversity may sometimes be appropriate (Hill et al. 2003), the data produced by ecologists surveying macro-organismal communities differ from data obtained by high-throughput sequencing of microbial communities in three key ways. First, in contrast to plant and animal assemblages, microbial assemblages are typically made up of more than one domain of life, thus necessitating the ability to quantify diversity across very disparate organism types. Second, many classical indices assume ecological communities are composed of unique species. However, traditional biological species concepts do not fit the natural histories of many microbial taxa that routinely undergo non-homologous recombination (Taylor et al. 2000, Rosselló-Mora and Amann 2001, Staley 2006) and sometimes lack sexual reproduction. (It is worth noting that the concept of species is widely questioned for macro-organisms as well (Mishler 2010).) Finally, unlike with macro-organisms, researchers are often unable to directly observe and characterize microbes and their traits *in situ* (Tiedje et al. 1999, Luo et al. 2007). The taxonomic/phylogenetic and functional genes of environmental microbes are now commonly sequenced, but it is still very difficult to link the taxonomy of an individual microbe to the environmental functions it carries out.

These differences create methodological issues when discrete, taxonomic-based metrics are used to analyze microbial community datasets. The culture-independent approaches employed by microbial ecologists usually survey a variety of genes, intergenic spacers, and transcripts, which are typically classified into discrete, taxonomic bins called Operational Taxonomic Units (OTUs). Homologous genetic fragments that share less than a certain percentage of nucleotide polymorphisms are classified as being in the same genus or species (e.g., 97% similarity of the 16S gene is widely uses for "species") (Horner-Devine et al. 2004, O'Brien et al. 2005, Buée et al. 2009). This cutoff fails to adequately include the homology (and thus shared ecological function) with which the species concept was originally conceived.

The limitations of applying traditional diversity indices to microbial datasets lacking clear species delineations leave a number of questions: How can we quantify diversity using methods that are better suited for microbial datasets which span multiple domains of life? Does including similarity in our analyses change our interpretation of patterns of microbial diversity? What is the utility of including multiple dimensions of microbial diversity (i.e., taxonomic and phylogenetic) in our analyses?

One promising new way to analyze microbial community diversity and address these questions is through the use of diversity profiles, which were recently developed by Leinster and Cobbold (2012) (Chao et al. 2010). These profiles are graphs that are used to display effective numbers of diversity (i.e., effective diversities). Effective diversities are mathematical generalizations of previous indices that behave much more intuitively, satisfying a number of desirable mathematical properties that provide meaningful percentage and ratio comparisons (Hill 1973). This is useful because many indices that have been traditionally used to describe macro-organismal community diversity and evenness can be quantitatively unintuitive (Inverse Simpson's Diversity Index, Shannon's Entropy, Gini-Simpson Index, etc.). For example, a community comprised of 10 hawks and 10 hummingbirds might experience a 50% decrease of both species, resulting in five hawks and five hummingbirds, but this change would not manifest as a 50% decrease in either Simpson Diversity or Shannon Diversity. Due to this, Hill (1973) and later Jost (2006) formulated effective number diversity metrics, which are simple entropies weighted by an order parameter, q. As the q parameter increases, the relative weight given to rare taxa in diversity index calculations declines. The effective diversity of order zero ($q = 0$) is

equivalent to species richness (the total number of entities), order 1 is proportional to the Shannon index, and $q = \infty$ is a measure of pure evenness (Leinster and Cobbold 2012).

Diversity profiles significantly improve these previous calculations of effective diversity by adding community similarity information into diversity calculations, using a similarity matrix, **Z**. The term "similarity" is used by Leinster and Cobbold to refer to the degree of distance or difference between organisms. The similarity matrix can accommodate genetic similarity, phenotypic similarity, or any other biologically meaningful source of similarity between two or more entities. Incorporating this information into similarity-sensitive calculations of community diversity can greatly alter conclusions regarding diversity levels (Leinster and Cobbold 2012). For example, when taking into account similarity between taxa, a bird community comprised of one hawk, one hummingbird, and one goose would be more diverse than a community of three distinct hummingbird species. However, if similarity between taxa were not taken into account, these communities would be classified as equally diverse.

For microbial communities, which are often characterized by phylogenetic molecular markers, the use of a metric based on the average evolutionary relatedness of a community conveys more information on the uniqueness and potential function of that community than does a discrete, OTU-based approach (Martiny et al. 2013). Recent work by Chao and colleagues (2010), which expands on research by Faith (1992), develops a measure of effective phylogenetic diversity. Effective phylogenetic diversity scales traditional diversity metrics by the hypothesized shared evolutionary history between taxa. Calculating phylogenetic diversity requires scaling raw taxonomic diversity by the shared evolutionary branches in a phylogeny. These branches can be either time-calibrated (ultrametric) or non-ultrametric. Even if a phylogeny is unavailable, the inclusion of cladistic data can be meaningful, if they accurately model shared ancestry within the study community. If the relative abundances of taxa or sequences are known, branches can also be weighted by abundance to compare the phylogenetic evenness among samples (Cadotte et al. 2010).

Given the differences between microbial and macro-organismal community data, the primary objective of this study was to evaluate the use of diversity profiles when analyzing microbial assemblages to determine whether the inclusion of similarity data (in our case, phylogenetic data) changes our interpretation of experimental and observational data. First, to explore whether diversity profiles alter our interpretation of microbial diversity data, we calculated diversity profiles for four datasets from different environments containing all domains of life and viruses. For comparison purposes, four statistics of pairwise community dissimilarity were calculated for the microbial datasets and plotted as dendrograms. Because diversity profiles can take into account the similarity of taxa and the relative importance of rare versus abundant taxa, we sought to evaluate how incorporating the phylogenetic similarity of taxa provides a different view of microbial diversity compared to traditional taxonomy-based metrics.

Second, we looked for evidence of bias and robustness of phylogenetic diversity profiles using simulated communities. We created numerous communities that varied in their rank abundance distributions, tree topologies, and whether ultrametric or non-ultrametric trees were used. Tree topologies were also simulated to create communities that spanned a large range of tree balances. Tree balance is determined by evolutionary processes, in particular lineage divergence and extinction rates and patterns, which differ greatly among real microbial communities (Mooers and Heard 1997). We wanted to compare how "naïve" diversity profiles (what Leinster and Cobbold (2012) term calculations that do not take taxa similarity information into account) and similarity-based diversity profiles are influenced by the topological

characteristics (e.g., tree ultrametricity, tree balance) of the sampled communities. We tested the concordance between taxonomic and phylogenetic measures of diversity and composition. We predicted that since OTU-based metrics are discrete transformations of phylogenetic measures, they would generally agree. Simulations (and real data) were also used to test whether this concordance is correlated with aspects of the sampled community including aspects of its phylogenetic topology, richness, and abundance distribution. Our analyses indicate that phylogenetic diversity profiles provide insights into microbial community diversity that would not be discernible with the use of traditional univariate diversity metrics.

Methods

Diversity profiles

Diversity profiles were calculated for experimental, observational, and simulated microbial communities, as presented in detail by Leinster and Cobbold (2012). Briefly, consider a fully sampled community that contains S unique species. The relative abundances of the species are calculated by $p_1, \ldots, p_s$, such that $p_i \geq 0$ and $\sum_{i=1}^{S} pi = 1$. Because $p_i \neq 0$, diversity profiles consider only species that are actually present in a community.

Information regarding the similarities between species in the community is taken into account by a matrix $\mathbf{Z} = (Z_{ij})$. The matrix has dimensions S X S, and Z_{ij} measures the similarity between the ith and the jth species. Similarity is scored such that $0 \leq Z_{ij} \leq 1$, so that 0 represents complete dissimilarity between two species and 1 represents identical species. When similarity information is not available, or authors do not wish to include it, $Z_{ij} = 1$ in all cases, and this results in a naïve calculation.

Diversity profiles were then calculated across the range of a sensitivity parameter, q, for the values of $0 \leq q \leq \infty$. At low values of q, such as $q = 0$, calculations of diversity are sensitive to rare taxa, and as q moves toward ∞, diversity calculations become more and more insensitive to the contributions of rare taxa.

For $q \neq 1, \infty$, the diversity profile calculation is thus ${}^{q}D^{\mathbf{Z}}(\mathbf{p}) = \left(\sum pi\left(\boldsymbol{Z}p\right)_i^{q-1}\right)^{\frac{1}{1-q}}$ where $\left(\boldsymbol{Z}p\right)_i = \sum_{j=1}^{S} \boldsymbol{Z}_{ij} p_j$. The resulting ${}^{q}D^{\mathbf{Z}}(\mathbf{p})$ is an effective number, and for certain values of q and $\mathbf{Z}$, ${}^{q}D^{\mathbf{Z}}(\mathbf{p})$ corresponds to a commonly used diversity index. For example, for naïve diversity profiles that do not take into account similarity between species, $q = 0$ is equivalent species richness, $q = 1$ is proportional to Shannon Diversity (Shannon 1948), $q = 2$ is proportional to $1/D$ (inverse Simpson Diversity) (Simpson 1949), and as q moves toward ∞, it is a measure of 1/Berger-Parker Evenness (Berger and Parker 1970).

We calculated diversity profiles for $0 \leq q \leq 5$. When plotting the profiles, we created larger insets for $1 \leq q \leq 2$ (Haegeman et al. 2013). For a more detailed description of the formulae used to calculate diversity profiles (e.g., their relationship to well-known diversity metrics, their potential benefits in diversity studies, examples of diversity profiles applied to macro-organism community datasets), refer to Leinster and Cobbold's work (2012).

Environmental microbial datasets

Diversity profiles were used to quantify the diversity of four microbial datasets obtained from different environments containing bacterial, archaeal, fungal, and viral communities. The

original four studies were conceived independently by co-authors of the current study, and we utilized these existing datasets to explore applications of diversity profiles to microbial community data. Providing complete details of each study is beyond the scope of the current study, but we have included brief descriptions of the studies' methods below, and the research questions and hypotheses that shaped the design of each study are detailed in Table 1. We have also provided predicted outcomes of each of the studies, based on data and hypotheses from the original studies (Table 2). For further details of each study, please refer to the publications cited below.

Acid mine drainage bacteria and archaea

Total RNA was purified from eight environmental biofilm communities, collected from the Richmond Mine at Iron Mountain, Northern California in 2010 and 2011. In addition, total RNA was extracted from five biofilms grown in laboratory bioreactors using Richmond Mine inoculum in 2009 and 2010. Biofilms were collected or harvested at varying stages of development, ranging from early (GS0), mid (GS1), and late (GS2), as described previously (Goltsman 2013).

RNA from all 13 samples was converted to cDNA and subject to Illumina library preparation and sequencing at the University of California Davis. Six environmental samples (from locations Env-1, Env-2, Env-3) and two bioreactor samples were sequenced using the HiSeq 2500 Illumina platform. Two environmental samples (from locations Env-2 and Env-4) and three bioreactor samples were sequenced using the GAIIx Illumina platform. A total of 256 million 75–100 bp long-reads were mapped to the small subunit (SSU) rRNA Silva database (including Archaea, Bacteria and Eukarya) with a similarity cutoff of 97% identity. SSU rRNA reads were then assembled using Cufflinks (Roberts et al. 2011), and clustered at 97% identity using uclust (Edgar 2010). SSU gene sequences were aligned using the SINA aligner webserver, and a phylogenetic tree was constructed using FastTree with options -gtr -nt -gamma. Normalized counts values obtained from Cufflinks were used as a measure of abundance of SSU rRNA genes sequences, as described earlier (Goltsman 2013).

Hypersaline lake viruses

As previously described in detail (Emerson et al. 2012, Emerson 2013), eight surface water samples were collected from two locations (A and B) within hypersaline Lake Tyrrell, Victoria, Australia (~330 g/L NaCl), with dates, locations, time scales, and sample IDs as follows: January 2007 (two samples, site A, two days apart, 2007At1, 2007At2), January 2009 (one sample, site B, 2009B), January 2010 (one sample, site A, 2010A; four samples, site B, each approximately one day apart, 2010Bt1, 2010Bt2, 2010Bt3, 2010Bt4). In the summer, when samples were collected, the lake dries and leaves residual briny "pools" in a few isolated sites. Sites A and B are different pools ~300 m apart.

Post-0.1 μm filtrates were concentrated via tangential flow filtration for the collection of viral particles, followed by DNA extraction and metagenomic sequencing. 454-Titanium technology (~400 bp reads) was used to sequence samples 2010Bt1 and 2010Bt3, and Illumina GAIIx paired-end technology (~100 bp reads) was used to sequence the remaining six samples, for a total of 6.4 billion bp. Previous analyses of these data show that there was no observable difference between the 454-Titanium data and the Illumina data (Emerson et al. 2012, Emerson et al. 2013a, Emerson et al. 2013b). Each sample was assembled separately via Newbler (Margulies et al. 2005), ABySS (Simpson et al. 2009), or Velvet (Zerbino et al. 2008). Genes

from all contigs >500 bp were predicted with Prodigal (Hyatt et al. 2010), and predicted genes longer than 300 bp were retained and clustered at 95% nucleotide identity, using uclust (Emerson et al. 2012). Corresponding predicted proteins were separately 1) annotated with InterProScan (Quevillon et al. 2005) and 2) clustered at 40% amino acid identity, using uclust (Emerson et al. 2012). In the absence of a universal marker gene, six viral "OTU groups" were chosen (Emerson et al. 2013b). Three were used for this study: methyltransferases (the most abundant annotation), concanavalin A-like glucanases/lectins (the most abundant annotation likely to be exclusive to viruses), and Cluster 667 (one of the largest protein clusters of unknown function). Proteins for each OTU group were aligned with MUSCLE (Edgar 2004), and a phylogenetic tree was constructed from the alignments, using FastTree (Price et al. 2009) with default parameters.

Subsurface bacteria

DNA was extracted from five sediment samples taken from *in situ* flow-through columns buried in sampling wells in a shallow, uranium and vanadium-contaminated aquifer in Rifle, Colorado as described previously (Yelton et al. 2013). Samples were from background sediment (B), sediment stimulated with carbon and vanadium addition (V1, V2), and sediment stimulated with carbon addition alone (A1, A2). Universal primers and gradient PCR were used to amplify the 16S small subunit ribosomal RNA gene from the organisms sampled.

HiSeq Illumina paired-end technology was used to sequence 2.7 megabases of PCR product at the University of California, Davis. The sequencing consisted of 26,954,412 100-base pair reads. Reads were mapped to reference sequences from the Silva database with the EMIRGE iterative algorithm (Miller at al. 2011, Miller et al. 2013). The genes were aligned to each other, using the SSU-align software (Nawrocki et al. 2009). The alignment was automatically masked with the ssu-mask program. Bacterial OTUs were then clustered at a 97% nucleotide identity cutoff, using usearch (Edgar 2010). A phylogenetic tree was constructed with the aligned sequences via the FastTree maximum likelihood method with options –gtr –nt and 1000 iterations of the FastTree bootstrap (Yelton et al. 2013).

Substrate-associated soil fungi

The goal of this study was to determine if substrate, space, time or plant community were the major determinants of fungal saprotrophic community composition. Sampling of buried substrates (straw and wood blocks) occurred on Bolinas Ridge on Mount Tamalpais in Marin County, California, USA along four 10 x 10 m blocks in 2007 and 2008, as previously described (Kerekes 2011). Two blocks were in the coastal grassland and two blocks were in the adjacent forest dominated by *Pseudotsuga menziesii*. The region is characterized as having a Mediterranean climate with a seasonal summer drought. DNA was extracted from 32 bait bags filled with sterile wheat straw and 32 small conifer wood blocks that had been buried (<10 cm) in both the grassland and forest blocks (16 straw samples and 16 wood samples were buried in each plant community type). Half of the straw and wood substrates were buried for six months (time point 1), while the others were buried for 18 months (time point 2).

DNA was purified, and the LSU region (LROR_F (Amend et al. 2010)/LR5-F (Tedersoo et al. 2008)) was PCR amplified with 10 bp MID barcodes. 454 Pyrosequencing 1/8 of a plate resulted in a total of 123,117 LSU sequences. Reads were trimmed and filtered using the QIIME software (Caporaso et al. 2010). Non-fungal taxa, sequences that resulted in no BLAST matches, and singletons were removed from the analysis. OTUs were conservatively determined at 95% sequence similarity. FastTree (Price et al. 2009) was used for phylogenetic tree building in

QIIME. For community analyses, only samples with at least 600 LSU sequence reads were included.

Analysis of datasets

Diversity profiles for each dataset were calculated using an R code adapted from Leinster and Cobbold (2012). For each community, both naïve diversity profiles and diversity profiles that took into account similarity information derived from the community phylogenies were calculated. The resulting profiles were then compared and analyzed. Specifically, we sought to identify differences between naïve and phylogenetic measures of diversity and community composition that would affect our interpretation of patterns in the data. The topology of the phylogenetic trees constructed from these datasets were quantified using Colless' I tree balance statistic (Colless 1982) with Yule normalization; high values of Colless' I correspond to imbalanced, asymmetric trees and low values correspond to more balanced trees (Table 3).

In order to compare the diversity calculations produced by diversity profiles to more traditional calculations of community composition for the same datasets, four different statistics of pairwise community dissimilarity were computed (abundance-weighted Jaccard, unweighted Jaccard, abundance-weighted UniFrac, and unweighted UniFrac). The Jaccard index, is the ratio of the number of taxa shared between two samples to the total number of taxa in each sample and then this ratio subtracted from one (Jaccard 1901). Pairwise phylogenetic dissimilarity for each sample was calculated using the UniFrac method (Lozupone and Knight 2005). This metric measures the proportion of unshared phylogenetic branch lengths between two samples. Ward's minimum-variance method (Ward 1963) was used to complete hierarchical clustering on the samples based on each dissimilarity metric and plot them as dendrograms. Please see Additional file 1 for these results.

Simulations

We simulated hundreds of microbial communities in order to better measure the degree to which differences between naïve and similarity-based diversity profiles are influenced by the abundance and phylogenetic distributions of microbial communities. Each simulated community was distributed according to one of four possible commonly fitted rank abundance distributions (Log Normal, Geometric, Log Series, or Uniform) and had a random phylogenetic tree topology. Tree topologies were simulated so as to create communities that spanned a large range of tree imbalances. Tree imbalance was quantified using Yule normalized Colless' I tree balance statistic (Colless 1982). Lastly, all trees were simulated in both ultrametric and non-ultrametric versions to test the effects of branch lengths on the diversity profiles.

To look for systematic differences between naïve and phylogenetic diversity profiles, we repeatedly (100 times) took a random sample of OTUs from two simulated communities and calculated the proportion of times that the naïve and phylogenetic diversity profiles agreed on which random sample was more diverse. We analyzed whether agreement between naïve and similarity-based diversity profiles systematically differed based on numbers of OTUs sampled, whether trees were ultrametric or non-ultrametric, Fisher's alpha diversity values, or tree imbalance values.

Results and Discussion

Given the potential limitations of applying traditional diversity indices to microbial datasets produced by high-throughput sequencing, we sought to evaluate microbial diversity using methods that might be better suited for microbial taxa that span multiple domains of life and multiple dimensions of diversity (e.g., taxonomic, phylogenetic). The advantages of using diversity profiles are that they encompass a number of other common diversity indices and allow for the incorporation of species similarity information.

We systematically tested diversity profiles as a metric for quantifying microbial diversity by analyzing four natural experimental and observational microbial datasets from varied environments that contained bacterial, archaeal, fungal, and viral communities. (Refer to Table 4 for summaries of these datasets.) For each of the four datasets, we specified plausible alternative hypotheses for the ecological drivers of each community's diversity (Table 1), as well as expected results (Table 2, Additional file 1: Table S1). Additionally, we tested diversity profiles on the simulated microbial datasets.

Naïve microbial diversity comparisons may vary with the sensitivity parameter, q

Diversity profiles calculated from the experimental and observational datasets provided insights into microbial community diversity that would not be perceivable through the use of a classical univariate diversity metric. The sensitivity of diversity profiles to rarity greatly affected diversity measurements. Richness calculations count all taxa equally, greatly overestimating the contribution of rare taxa to diversity, whereas diversity measurements at high values of q are insensitive to the contribution of rare OTUs. Diversity profiles illustrate this stark contrast and highlight the question of the importance of ultra-rare taxa, the "rare biosphere" of Sogin et al. (2006). Previously, these ultra-rare taxa were not included in diversity calculations because they were not detected using older methods of measuring microbial taxa (clone libraries, low depth sequencing, DGGE, etc.). Newer techniques such as deep short-read sequencing have revealed the existence of these taxa, but introduced more bias into older diversity indices such as species richness calculations. The datasets analyzed here demonstrate the importance of rare taxa.

This is clearly indicated by the viral data from the hypersaline lake viruses dataset. For the viral gene clusters described in this study, there was some disagreement in the relative diversity rankings of samples across the range of q plotted in all three naïve diversity profiles (Table 1, Figure 1, Additional file 1: Figures S2, S3). First, if diversity of the putative genes falling under Cluster 667 were analyzed with the naïve analysis using only species richness (q = 0 in the diversity profile), the resulting calculations would have indicated that the 2009B sample was the most diverse (Figure 1). However, by q = 1 (which is proportional to calculating Shannon index) and for all higher values of q, the sample 2009B had the lowest diversity within the dataset. This change in ranking at higher values of q indicates that the 2009B sample had many rare taxa, because as q increases, the weight given to rare taxa in diversity profile calculations decreases (Leinster and Cobbold 2012). Secondly, in the naïve diversity profile for the putative methyltransferase group, the lines representing the diversity of the 2007A, 2009B, and 2010B samples crossed each other numerous times between q = 0 and q = 5 (Additional file 1: Figure S2). Lastly, in the naïve profile for the putative concanavalin A-like glucanases/lectins group, the 2010B samples were as diverse as or more diverse than the 2007A samples at q = 0, but the diversity of 2010B samples dropped sharply and remained lower than all other samples after approximately q = 0.5 (Additional file 1: Figure S3). In the case of viral diversity, ultra-rare

taxa play an important role in rapid evolution to allow new viruses to infect hosts that are constantly evolving defense mechanisms. Thus, diversity calculated at low values of q, which are sensitive to rare taxa, is the more appropriate measure of viral diversity.

We see similar results for the acid mine drainage dataset. At $q = 0$ (species richness) in the naïve analysis, the Env-3 at growth stage 2 sample is the most diverse sample, but the sample's diversity decreases and is surpassed by the growth stage 0 bioreactor sample and both Env-1 samples between $q = 1$ and $q = 2$ (Figure 2), demonstrating that the bioreactor and Env-1 samples were less even than the Env-3 sample at growth stage 2. Thus, for this dataset as well as for the hypersaline lake viruses dataset, evaluating the diversity of the microbial communities at multiple values of q leads to a different interpretation of the results and response to the original hypotheses (Table 1).

Diversity profiles do not always add new information to analyses of natural microbial datasets. In some cases, such as with the naïve profiles of the subsurface bacteria dataset, the most diverse samples in a dataset were always calculated as the most diverse, across the entire range of q in the naïve profile (Figure 3). Thus, whether we quantified diversity using species richness, Shannon diversity, or diversity profiles, we would arrive at the same result. In general, our findings provide evidence for the utility of diversity profiles to analyze microbial datasets, even when similarity information is not taken into account, because they allow researchers to visualize multiple diversity indices across the range of q in the same place after just one calculation. They also clearly provide information about the effects of rare species in a sample on diversity calculations.

Similarity information may alter microbial diversity calculations

The analyses presented here demonstrate the value of using diversity profiles to incorporate phylogenetic diversity as a measure of taxa similarity into diversity calculations. For all four microbial datasets we analyzed, we saw key distinctions between naïve taxonomic diversity calculations and those that incorporated phylogenetic information. For example, in the subsurface bacterial dataset, naïve measurements of OTU richness for each treatment indicated that the background sample (no treatment) contained the highest diversity for all values of q (Table 2, Figure 3A). Additionally, naïve measurements of both acetate-only samples were more diverse than the samples amended with both acetate and vanadium. These were the expected results as the experiment involved a treatment that should have selected for taxa that could use acetate as a carbon source and vanadium as an energy source (Table 1).

Phylogenetic results, on the other hand, suggested that the vanadium-acetate samples were as diverse as background samples and more diverse than the acetate-only treatments (Table 2, Figure 3B), indicating that perhaps the ability to use vanadium for energy or to tolerate its presence was more phylogenetically widespread than expected. Previous analysis of these data using Faith's phylogenetic diversity metric found the background sediment to be most phylogenetically diverse (Yelton et al. 2013), which Figure 3B also shows at $q = 0$. However, the crossing of the background sample and the acetate and vanadium treated samples when $1 \leq q \leq 2$ in Figure 3B indicates a greater diversity of common taxa in the treated sites. This indicates that adding abundance information to measures of phylogenetic diversity through the use of diversity profiles can add depth to the interpretation of diversity calculations.

In another example, in forest samples at T = 1 in the substrate-associated soil fungi dataset, wood substrates contained greater naïve taxonomic diversity. This higher diversity on wood substrates compared to straw substrates was hypothesized because the wood substrate is

more complex and requires a larger group of fungi to decompose it compared with a simpler substrate, such as straw (Table 1). However, the wood substrates actually contained lower phylogenetic diversity than straw substrates (Additional file 1: Figure S4). These results indicate that the fungal communities growing on wood substrates contained more member taxa that were closely related to each other, because when phylogenetic similarity was included in diversity calculations, the diversity of wood substrate fungal communities decreased.

Similarly, when analyzing the grassland samples of the substrate-associated soil fungi dataset, the wood substrate samples contained greater naïve taxonomic diversity at both time points than the straw substrates (again, as hypothesized in Table 1), within the range of $0 \leq q \leq 5$ (Figure 4A). However, when phylogenetic similarity was included, the fungi growing on straw substrates at T = 1 were more diverse than the fungi growing on wood substrates at T = 1, within the range of $1 \leq q \leq 5$ (Figure 4B). This indicates that the fungal communities growing on straw substrates in the grassland at T = 1 contained taxa that were less closely related to each other (more phylogenetically diverse) than the taxa growing on wood substrates at T = 1, because when phylogenetic similarity was considered, the diversity of straw substrate fungal communities increased. There was also considerable overlap and crossing in the phylogenetic diversity profile between $1 \leq q \leq 3$, which was not apparent in the taxonomic profile.

This demonstrated capacity of diversity profiles to incorporate effective phylogenetic diversity, as well as other measures of similarity between taxa, is particularly meaningful for analyzing microbial diversity data. Macro-organismal ecologists have long been concerned with the interactions between an organism's traits and aspects of its ecology, such as its niche axes or its role in ecosystem processes (Hooper and Vitousek 1997, Tilman et al. 1997, Silvertown 2004, Ackerly and Cornwell 2007). Many macro-eukaryote traits, when mapped to phylogenies, show evidence for phylogenetic conservatism (Chazdon et al. 2003, Brumfield et al. 2007). That is, certain traits are shared more often by closely related taxa than would be expected by chance. Even bacteria and archaea show evidence for trait conservatism, despite the role of non-homologous recombination in their evolutionary history (Placella et al. 2012). This implies that the phylogenetic distribution of a microbial assemblage can, thus, influence ecosystem processes via differences in the suite of traits present. Phylogenetic trait conservatism in microbes also has practical implications, such as potentially guiding current research in drug discovery or biodegradation (Galvão et al. 2005, Ferrer et al. 2009, Singh and Macdonald 2010).

Diversity analyses of environmental microbial samples can span all domains of life. It is thus highly desirable to evaluate and critically assess a method that can address the diversity of a microbial assemblages effectively across domains, as well as across samples with substantial differences in rare membership, while using a full complement of the information contained in DNA and RNA sequence analysis. As there is no universal marker gene for viruses, there are no robust means of determining viral phylogeny from community sequencing data. Apart from a few groups of well-characterized viruses, it is difficult to characterize viral phylogenetic relationships at all. In our similarity-based profiles, we assume that sequence and, therefore, tree similarity are proxies for phylogenetic similarity. This is reasonable for phylogenetically informative genes, such as the SSU rRNA genes in cellular organisms. However, in the case of genes from the hypersaline virus dataset, and any other viral metagenomic data to which diversity profiles may be applied, this is almost certainly not true. In our application of sequence similarity-based diversity profiles to viruses, we essentially (incorrectly) inferred phylogeny from functional genes that are likely subject to extensive horizontal gene transfer. While these genes are still informative in that they might correspond to the host range and thus the viruses'

community function, we suggest that naïve diversity profiles will be more useful for analyses of viral assemblages than similarity-based profiles, unless a more robust means of determining viral phylogeny is discovered.

Diversity profile simulations

The four microbial datasets analyzed in this study were well-suited to test the application of diversity profiles to microbial data, particularly because they spanned multiple domains of life and dimensions of diversity. However, while treatment replicates were included in the diversity profiles for two of the datasets (hypersaline lake viruses, subsurface bacteria dataset), they were not included for the other two datasets. Therefore, statistical tests were not performed to determine whether the diversity of a group of samples was significantly higher or lower than other groups. Additionally, while it is noteworthy that we analyzed four unique microbial datasets within this study, our conclusions of how diversity profiles perform when analyzing microbial data were limited based on this relatively small number of datasets.

In order to address these shortcomings of the data, we simulated microbial communities. Simulations allowed us to utilize diversity profiles at the scale of hundreds of simulated microbial datasets with a range of abundance distributions and phylogenetic tree topologies, so that analyses were carried out with greatly increased replication. The major finding from this simulation study is that when we repeatedly took a random sample of OTUs from two simulated communities and compared their diversity, naïve and similarity-based diversity profiles agreed only approximately 50% of the time in their classification of which sample was most diverse (95% confidence interval was 29.8% to 74.6%, mean was 52.2% across all experiments). This finding is a strong argument for analyzing more than taxonomic diversity when quantifying the diversity of microbial communities. The evolutionary or phylogenetic distance among members of microbial consortia is arguably foundational in assessing diversity of these nodes of life that span the domains. It appears that microbial diversity analyses should include similarity information whenever it is available or its omission should be appropriately justified. Such similarity information need not include continuous evolutionary distances, but could be as simple as assigning similarity values based on general taxonomic group.

Our simulations showed that, to some extent, the choice of q did effect the agreement between naïve and similarity-based diversity calculations. Generally speaking, for small positive q values it appears that there was greater agreement between naïve and similarity-based diversity calculations. These differences were statistically significant when the difference in proportion of agreement between two q was ~ 0.15 (based on Z test for two population proportions). Turning to the impacts of tree typology and sample relative abundance distributions, our results showed that the percent agreement between the naïve and similarity-based diversity calculations decreased slightly with increasing skewed abundance distributions (Figure 5C) and increasing tree imbalance (Figure 5D). This finding is significant because, while tree shape changes greatly between different sized trees (Blum and François 2006), skewed abundance distributions (Fisher et al. 1943, Magurran and Henderson 2003) and higher tree imbalances (Simpson 1949, Blum and François 2006) are likely better representations of the majority of true environmental communities than perfectly balanced abundance distributions and phylogenies would be. In contrast, the percent of agreement increased slightly with increasing sample size (Figure 5A) and the use of non-ultrametric trees (Figure 5B), which are also likely good representations of the majority of true environmental microbial communities that may include thousands of OTUs (e.g., Sunagawa et al. 2010) and may produce undated non-ultrametric trees. Since these

simulations of phylogenetic trees with characteristics that resemble those of real datasets showed both slight increases and decreases in the percent agreement between the naïve and similarity-based diversity calculations, the percent agreement between naïve and similarity-based diversity calculations for real datasets is probably approximately 50%.

Conclusion

This study explored whether similarity-based diversity profiles can aid our interpretation of microbial diversity. The findings indicate that the use of phylogenetic metrics and effective numbers can provide additional insight into the diversity of microbial communities when combined with naïve analyses that do not take into account similarity information or multiple diversity metrics. The ongoing question of how to best analyze microbial community datasets is paramount to deducing the processes that affect the composition and function of microbial communities. The type of information and metric used to measure biological diversity in any study of microbial diversity is a decision that must be well-justified prior to hypothesis testing instead of being made arbitrarily based solely on which metrics are popularly used by plant and animal ecologists. This justification, in turn, should be based on evidence produced by work, such as this study, that has systematically tested the efficacy and utility of these diversity metrics under a range of situations.

Availability of Supporting Data

The R code adapted from Leinster and Cobbold [17] and used to calculated diversity profiles is available for download and use at https://gist.github.com/darmitage. The hypersaline lake viruses raw sequencing reads are available in the NCBI BioProject (accession number PRJNA81851, http://www.ncbi.nlm.nih.gov/bioproject/?term=PRJNA81851). The subsurface bacteria dataset is available at: http://banfieldlab.berkeley.edu/SOM/yelton2012/.

Tables

Table 1. Research questions and hypotheses that shaped the design of the four environmental microbial community datasets

	Research Questions	***Hypotheses***
Acid mine drainage bacteria and archaea	1) Are environmental (Env) samples more diverse than bioreactor (BR) biofilms?	H1: Bioreactor growth conditions usually have a higher pH than the environment, and the geochemistry of the drainage might differ from growth media. Thus, environmental biofilms are expected to be more diverse than bioreactor-grown biofilms.
	2) Is biofilm diversity higher at higher stages of biofilm development?	H2: As biofilms begin to establish, early growth-stage biofilms are expected to be less diverse. As they mature, more organisms join the community, increasing diversity.
Hypersaline lake viruses	1) How do viral diversities change across spatiotemporal replicates?	H1: Viral diversity will be greatest in pools with larger volume (2010A and 2007A samples). H2: Community dissimilarity will cluster by site, then by year.
Subsurface bacteria	1) Does acetate addition affect the diversity and composition of soil microbial communities?	H1: Acetate addition will stimulate growth of a subset of the microbial community capable of using it as an electron donor.
	2) Does vanadium addition affect the diversity and composition of soil microbial communities?	H2: Vanadium addition will reduce the diversity and evenness of the communities and favor those who can both use acetate as an electron donor and vanadium as an electron receptor and/or tolerate vanadium at high concentrations.
Substrate-associated soil fungi	1) How do plant community type (forest vs. grassland), substrate type (wood vs. straw), and time (6 months vs. 18 months) affect saprotrophic fungal assemblages?	H1: Wood substrates will be more diverse than straw substrates, because the wood substrate is more complex and requires a larger group of fungi to decompose it compared with a simpler substrate, such as straw. H2: Plant community type will have a greater effect on diversity than substrate type or time, because it will determine which fungi can colonize a substrate.

Table 2. Results of the diversity profiles for the four environmental microbial community datasets

	Treatment	*Naïve Profiles Results*	*Was This Predicted?*	*Similarity Profiles Results*	*Was This Predicted?*
Acid mine drainage bacteria and archaea	HiSeq	BR less diverse than most Env. samples	Yes	BR less diverse than Env. samples	Yes
		High GS only more diverse than early GS for Env-1	No	Highest GS (GS 2) is most diverse of all samples	Yes
	GAIIx	BR more diverse than Env-2, but less than Env-4	No	Env. samples mostly more diverse than BR	Yes
		Higher GS is less diverse than lower GS for BR	No	Highest GS is most diverse of all samples	Yes
Hypersaline lake viruses	N/A	Diversity greater in larger pools	Yes (2010A for 2/3 genes; not true for Cluster 667)	Diversity greater in combined 2007A samples and/or 2010A	Yes
Subsurface bacteria	N/A	Background > Acetate > Vanadium + acetate	Yes	Background ≈ Vanadium + acetate > Acetate	No
Substrate-associated soil fungi	Grassland	At all q: Wood T2 > Wood T1 > Straw T1 > Straw T2; No crossing along *q*	Yes	Straw T2 least diverse at all *q*	Yes
				At *q* = 0, Straw T1 has second lowest diversity, but by *q* = 3, has highest diversity	No
				Wood T2 > Wood T1 at all *q*	Yes
	Forest	At all *q*: Wood T1 > Straw T1 > Wood T2 > Straw T2; No crossing along *q*	No	At all *q*: Straw T1 > Wood T1 > Wood T2 > Straw T2; No crossing along *q*	No

Table 3. Yule normalized Colless' I tree balance calculations for the four environmental microbial community datasets

	Number of Tips	Yule Normalized Colless' I
Acid mine drainage bacteria and archaea	158	5.27
Hypersaline lake viruses: Cluster 667	71	0.33
Subsurface bacteria	10405	34.85
Substrate-associated soil fungi	1973	9.81

Table 4. Summaries of the four environmental microbial community datasets

	Dataset Summary	***Resulting Data***
Acid mine drainage bacteria and archaea	Total RNA was collected from 8 environmental biofilms and 5 bioreactor biofilms at varying stages of development: early (GS0), mid (GS1), and late (GS2). RNA from all samples was converted to cDNA. 6 environmental and 2 bioreactor samples were sequenced using HiSeq 2500 Illumina. 2 environmental and 3 bioreactor samples were sequenced using GAIIx Illumina.	159 SSU-rRNA sequence fragments were identified in 13 biofilms. The number of reads and SSU-rRNA sequences assembled from the GAIIx and the HiSeq platforms differed greatly; thus the rarefied data from these sequencing methods were analyzed separately (HiSeq: Figure 2, GAIIx: Additional file 1: Figure S1).
Hypersaline lake viruses	8 surface water samples were collected within a hypersaline lake as follows: Jan. 2007 (2 samples, site A, 2 days apart, 2007At1, 2007At2), Jan. 2009 (1 sample, site B, 2009B), Jan. 2010 (1 sample, site A, 2010A; 4 samples, site B, each ~1 day apart, 2010Bt1, 2010Bt2, 2010Bt3, 2010Bt4). 454-Titanium was used to sequence samples 2010Bt1 and 2010Bt3. Illumina GAIIx was used to sequence the remaining 6 samples.	630 methyltransferase genes, 411 concanavalin A-like glucanases/lectins, and 71 putative genes falling under Cluster 667 were assembled from the viral metagenomic reads (Methyltransferase: Additional file 1: Figure S2, Concanavalin: Additional file 1: Figure S3, Cluster 667: Figure 1).
Subsurface bacteria	DNA was extracted from 5 sediment samples taken from *in situ* flow-through columns buried in sampling wells in a shallow, uranium and vanadium-contaminated aquifer: background sediment (B), sediment stimulated with carbon and vanadium addition (V1, V2), and sediment stimulated with carbon addition alone (A1, A2). HiSeq Illumina was used to sequence 16S SSU-rRNA PCR product.	25,966 OTUs were identified from 5 subsurface samples (Figure 3).
Substrate-associated soil fungi	DNA was extracted from 32 straw bait bags and 32 wood blocks that were buried in grassland and forest (16 straw and 16 wood in each). Half of the substrates were buried for six months (time point 1) and half for 18 months (time point 2). 454-Titanium was used to sequence the PCR amplified LSU region.	508 total OTUs were identified within all substrate samples (Grassland: Figure 4, Forest: Additional file 1: Figure S4).

Figures

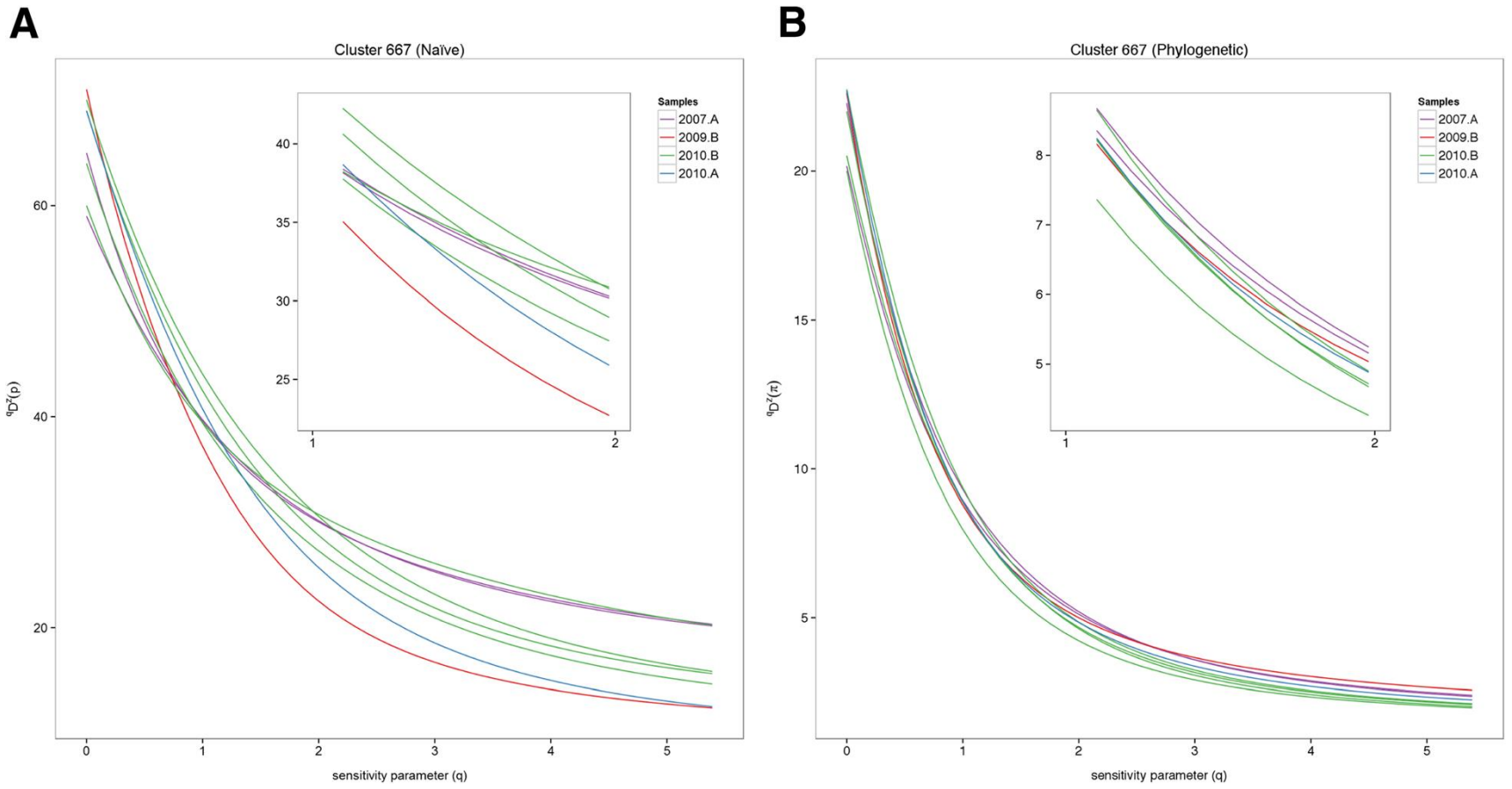


Figure 1. Hypersaline lake viruses Cluster 667 diversity profiles. (A) Naïve and (B) similarity-based (phylogenetic relatedness) diversity profiles calculated for Cluster 667 from the hypersaline lake viruses data.

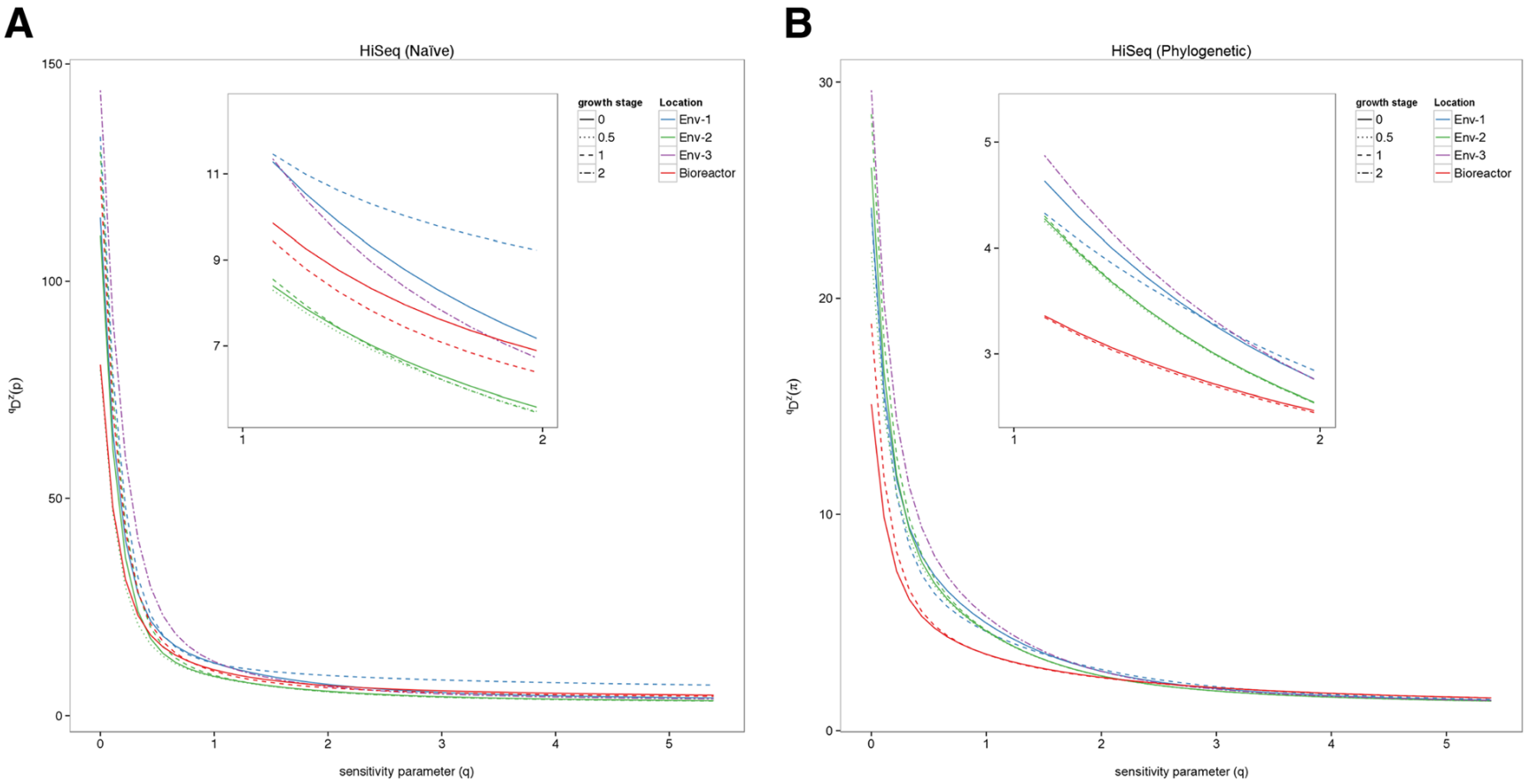


Figure 2. Acid mine drainage bacteria and archaea (HiSeq) diversity profiles. (A) Naïve and (B) similarity-based (phylogenetic relatedness) diversity profiles calculated from the acid mine drainage bacteria and archaea HiSeq data.

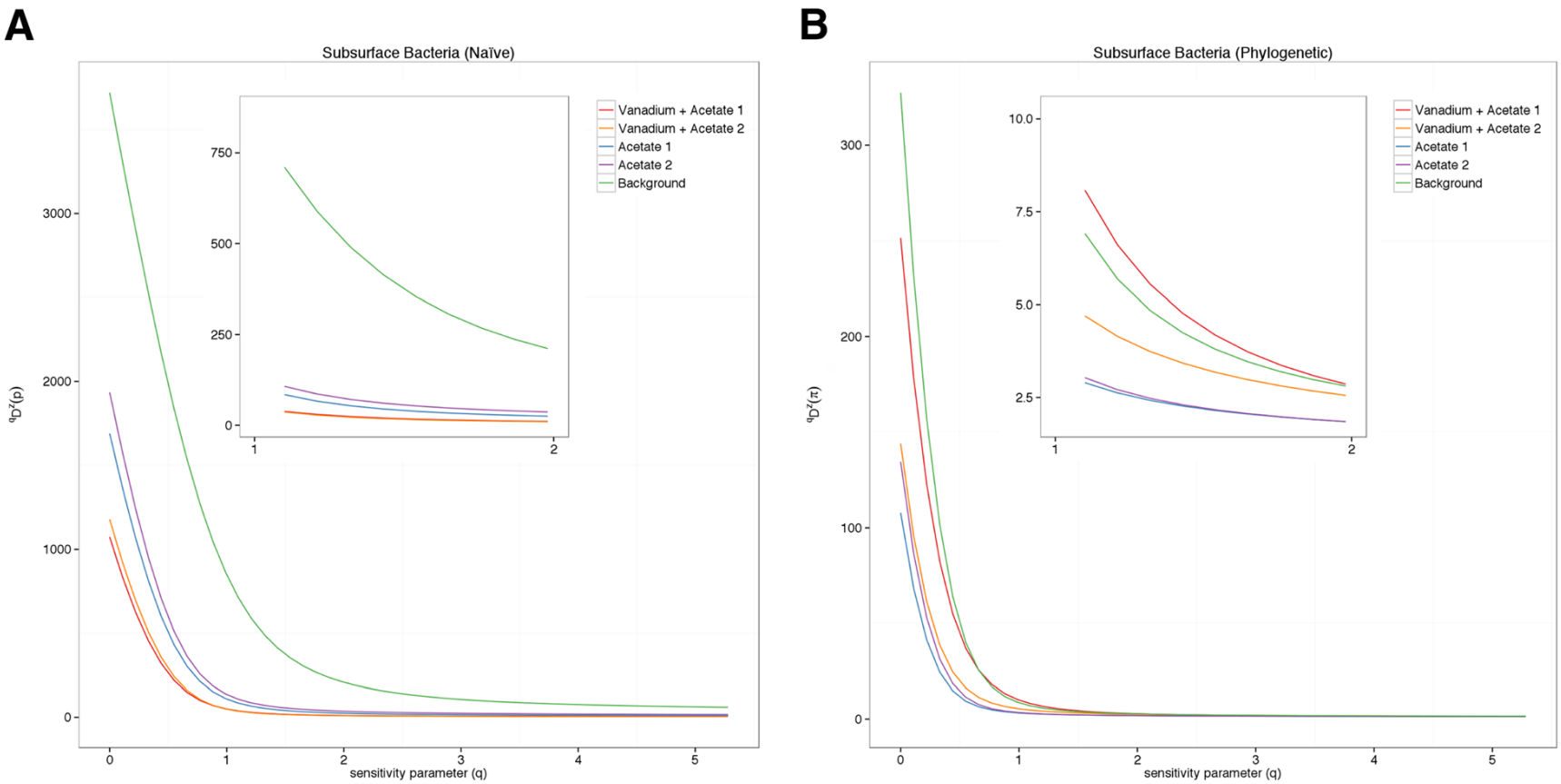


Figure 3. Subsurface bacteria diversity profiles. (A) Naïve and (B) similarity-based (phylogenetic relatedness) diversity profiles calculated from the subsurface bacteria data.

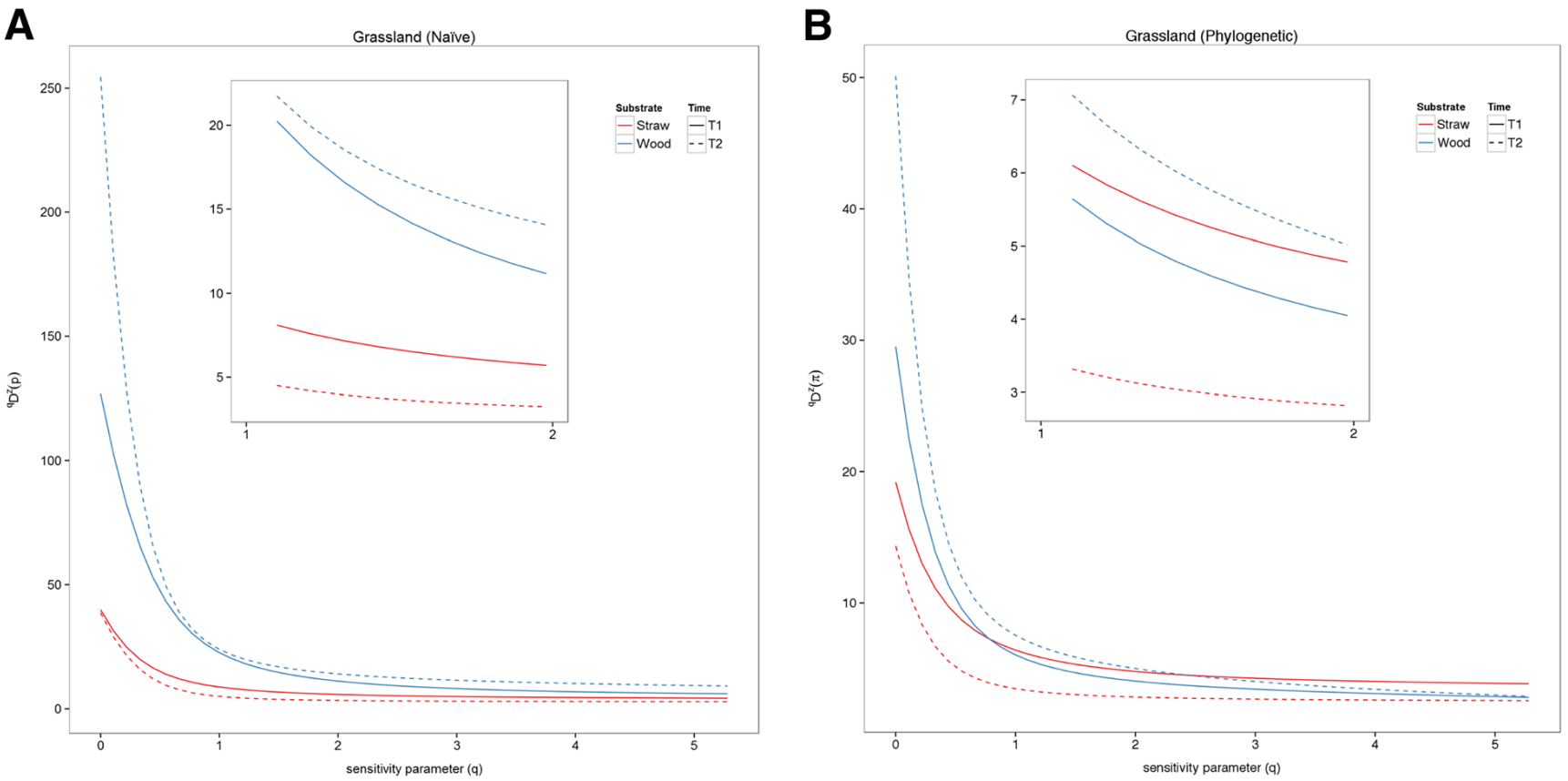


Figure 4. Substrate-associated soil fungi grassland diversity profiles. (A) Naïve and (B) similarity-based (phylogenetic relatedness) diversity profiles calculated from the substrate-associated soil fungi grassland data.

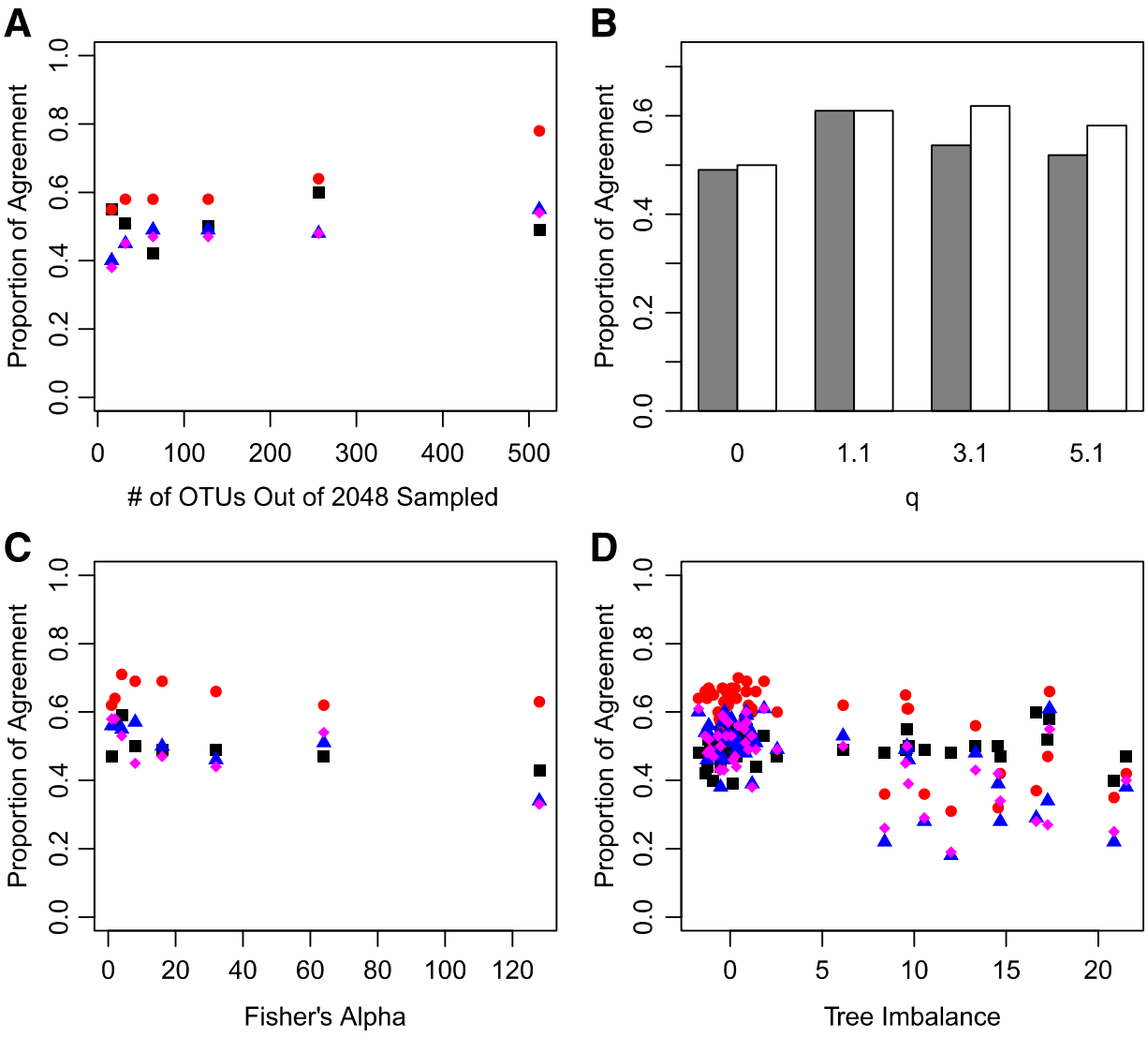


Figure 5. Agreement between naïve and similarity-based diversity profiles for different simulated communities. (A) For different numbers of OTUs sampled from the total pool of 2048, (B) for ultrametric (grey) and non-ultrametric trees (white), (C) for communities with different Fisher's alpha diversity values, (D) for communities with different tree imbalances. For panels (B), (C), & (D) sampled communities sized was 256; (A), (B), & (C) tree imbalance was 9.54; (A), (B), & (D) community abundance distribution was logseries with a Fisher's Alpha of 1. Proportion of agreement is based on 100 simulations. "black square symbol" ($q = 0$), "red circle symbol" ($q = 1.1$) "blue triangle symbol" ($q = 3.1$), "magenta triangle symbol" ($q = 5.1$).

Supplementary Material

Community dissimilarity comparisons

In order to compare the diversity calculations produced by diversity profiles to more traditional calculations of community composition for the same datasets, four different statistics of pairwise community dissimilarity were computed (abundance-weighted Jaccard, unweighted Jaccard, abundance-weighted UniFrac, and unweighted UniFrac). Please see the Methods section of the manuscript for further description of these indices.

Acid mine drainage bacteria and archaea

Data from the HiSeq-platform showed very similar clustering topologies between Jaccard and UniFrac, as well as between abundance-weighted and unweighted samples (S1, Figure S5). In the weighted calculations, the bioreactor samples, which were taken from the same reactor at different time points, matched more closely with each other than with environmental samples. However, samples clustered rather randomly in the unweighted calculations. The topologies of hierarchical clustering results were also similar between the Jaccard and UniFrac methods for both abundance-weighted and unweighted GAIIx samples: The bioreactor samples cluster together in weighted calculations (Figure S6).

Hypersaline lake viruses

The relative topological differences between community dissimilarity measures (UniFrac and Jaccard) for the hypersaline lake viruses dataset were greater between abundance-weighted and unweighted samples of the same type of information (phylogenetic and taxonomic) than between the two metrics under the same abundance-weighting assumption (Table S1, Figures S7, S8, S9).

Subsurface bacteria

Hierarchical clustering of phylogenetic (UniFrac) and taxonomic (Jaccard) community dissimilarity indices gave similar topologies (Table S1, Figure S10). However, these topologies differed between abundance-weighted and presence/absence formulations. In the latter, the background and acetate samples were most similar. When weighted by sequence abundances the vanadium plus acetate treatments clustered more closely with acetate-only treatments.

Substrate-associated soil fungi

The topology of hierarchical dissimilarity clusters were most similar between unweighted Jaccard and UniFrac methods (Table S1, Figure S11). The abundance-weighted variants of these methods arrived at slightly different topologies. However, both weighted and unweighted analyses grouped samples from similar habitats together, though the clustering of substrates or time points for each habitat varied with the method used.

Table S1. Results of the community composition analyses (Jaccard and Unifrac) for the four environmental microbial community datasets.

	Naïve Composition Results	*Was This Predicted?*	*Similarity Composition Results*	*Was This Predicted?*
Acid mine drainage bacteria and archaea	Bioreactors cluster separately from environmental samples	No	Bioreactors cluster separately from environmental samples	Yes
Hypersaline lake viruses	Clusters by site then year	Yes	Clusters by site then year	Yes
Subsurface bacteria	Clusters reflect treatments	Yes	Clusters reflect treatments	Yes
Substrate-associated soil fungi	Unifrac and Jaccard mostly cluster first by Community	Yes	Unifrac clusters by Substrate+Community, then by Community	Yes
	They then cluster alternatively by either Timepoint or Substrate	No	Jaccard mostly clusters like Unifrac, except Forest samples cluster first by Community+Timepoint (not substrate)	No
	Straw in Grassland T2 and Straw in Forest T2 break the above trends and cluster together in both Unifrac and Jaccard	No		

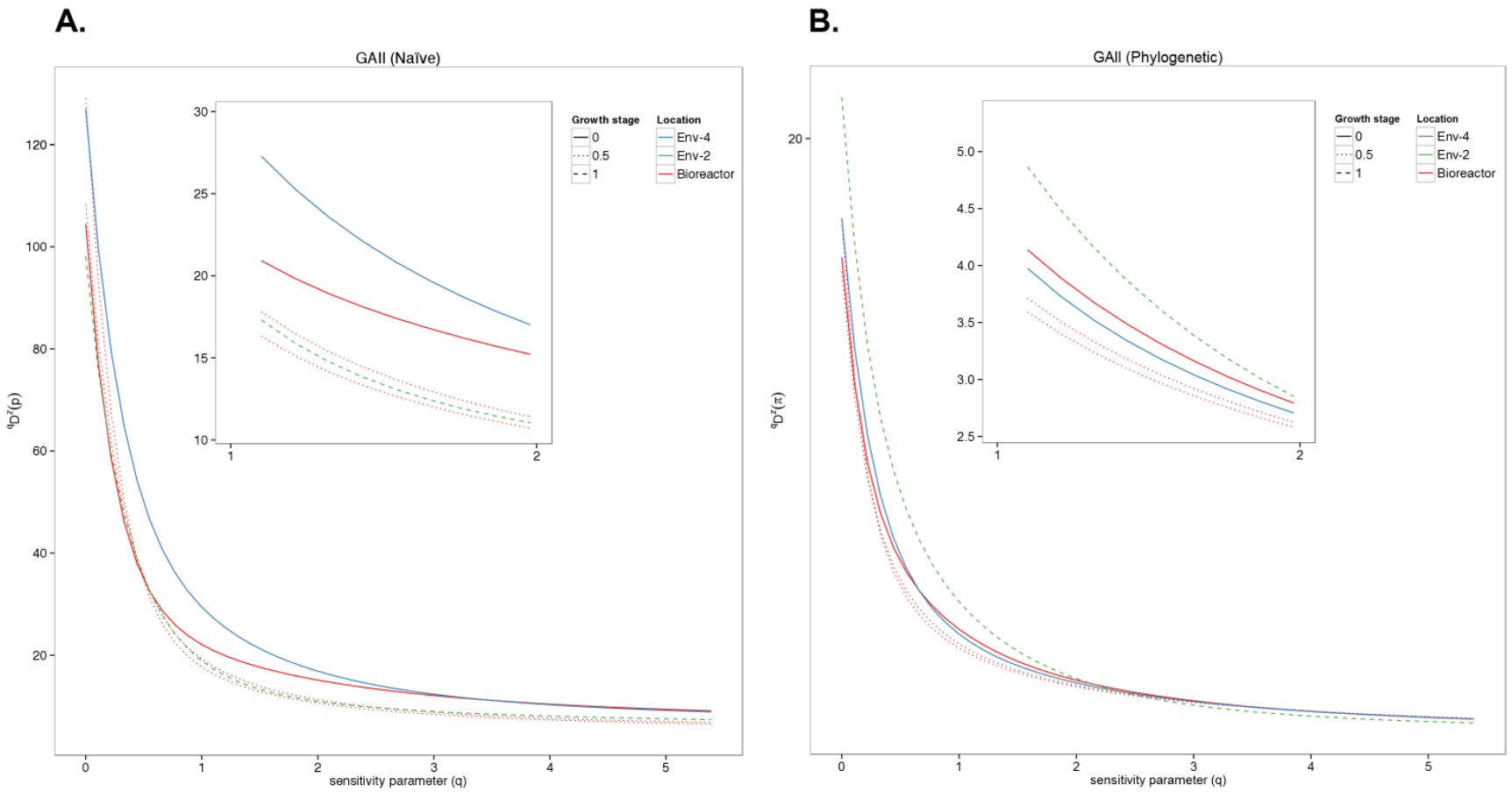


Figure S1. Acid mine drainage bacteria and archaea (GAIIx) diversity profiles. (A) Naïve and (B) similarity-based (phylogenetic relatedness) diversity profiles calculated from the acid mine drainage bacteria and archaea GAIIx data.

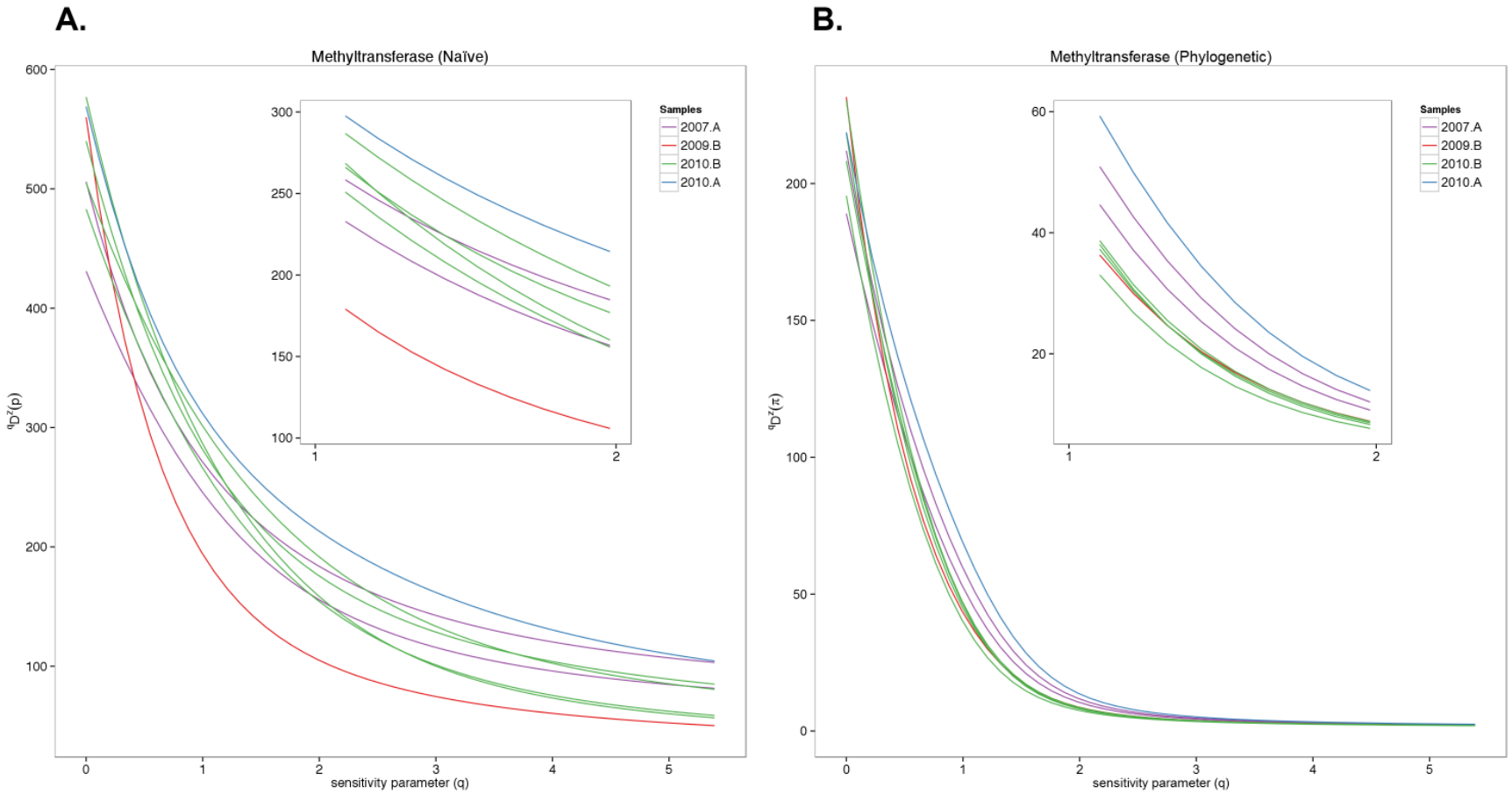


Figure S2. Hypersaline lake viruses methyltransferase diversity profiles. (A) Naïve and (B) similarity-based (phylogenetic relatedness) diversity profiles calculated from the hypersaline lake viruses methyltransferase data.

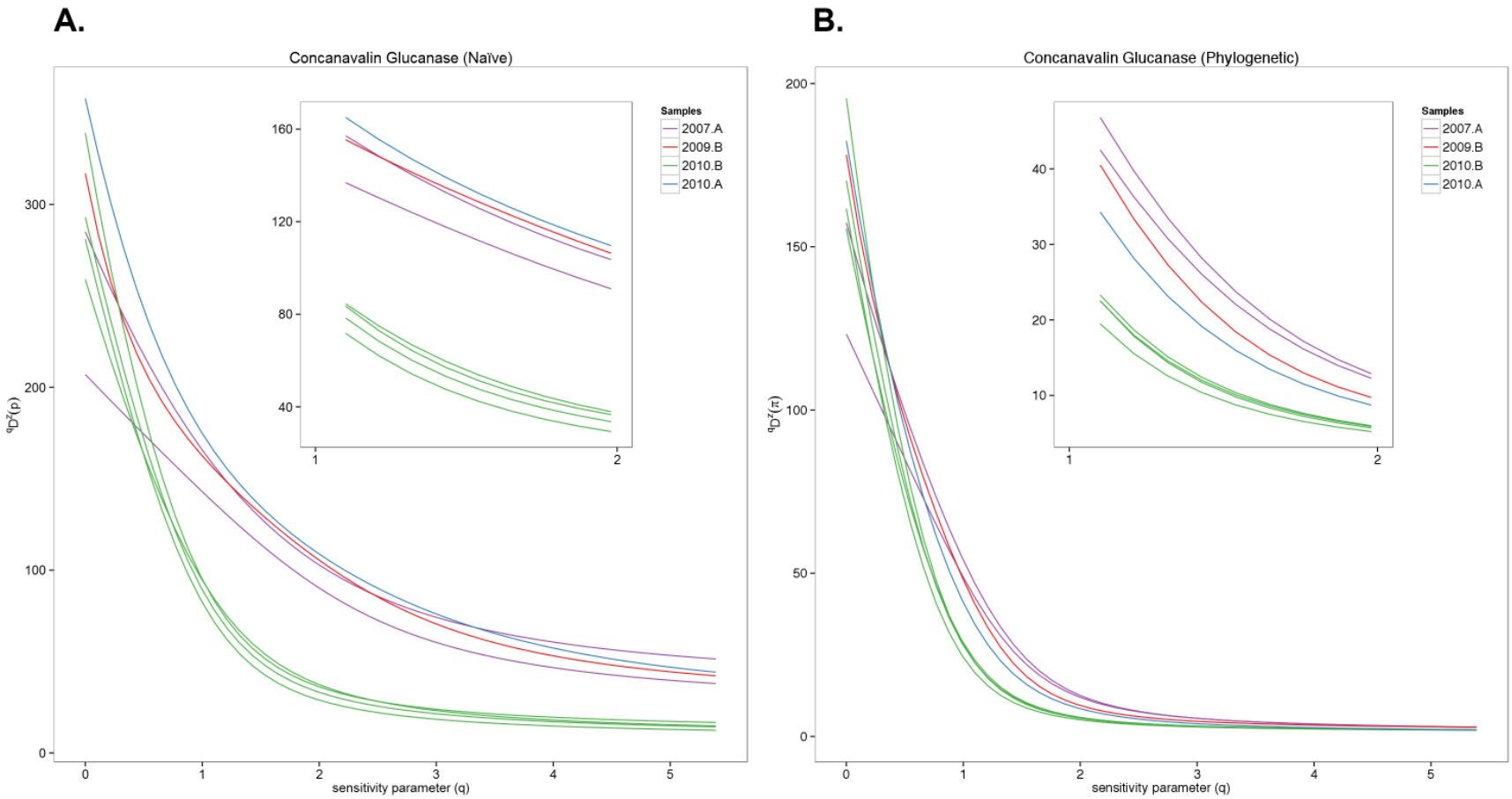


Figure S3. Hypersaline lake viruses concanavalin A-like glucanases/lectins diversity profiles. (A) Naïve and (B) similarity-based (phylogenetic relatedness) diversity profiles calculated from the hypersaline lake viruses concanavalin A-like glucanases/lectins data.

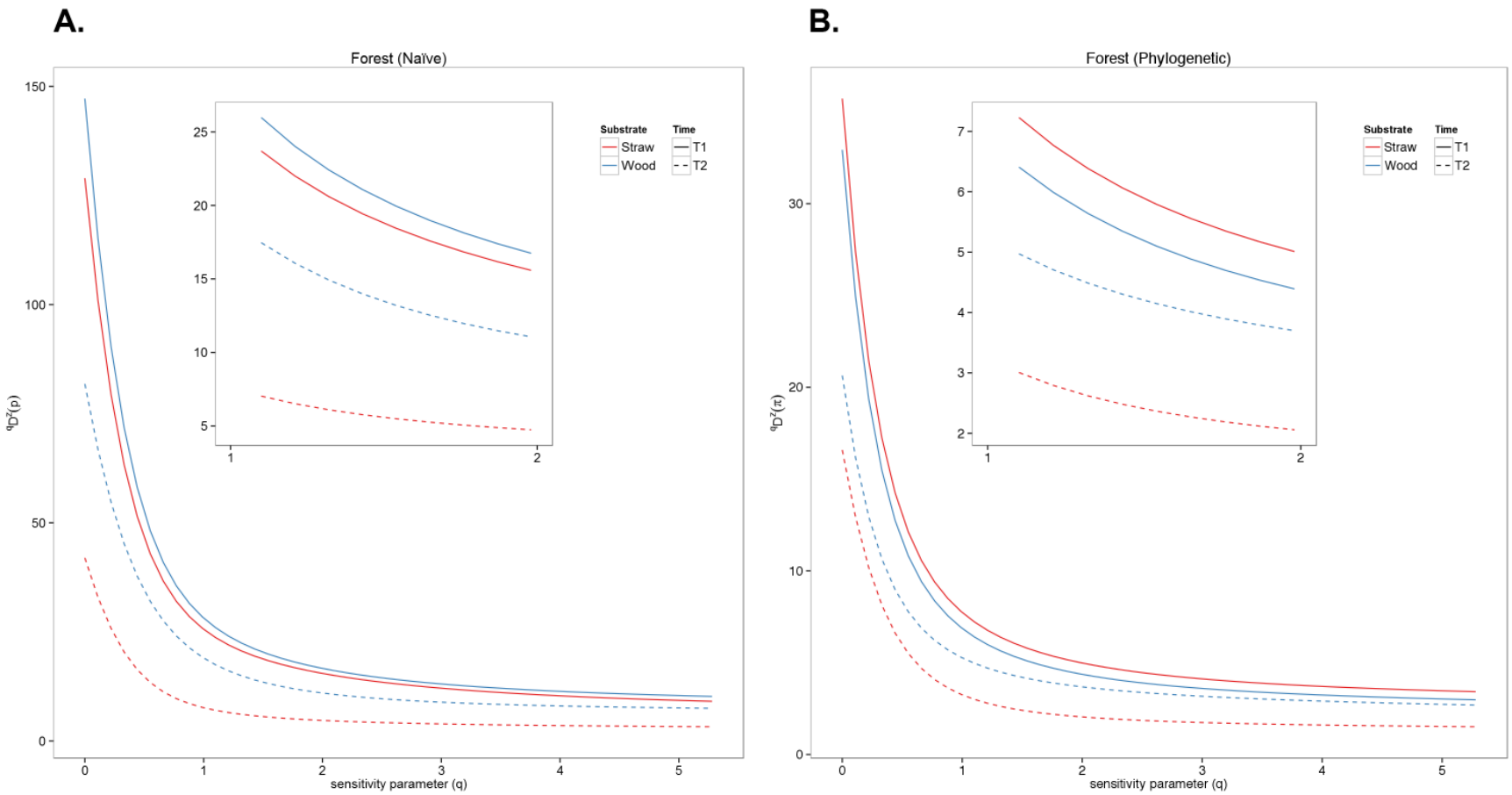


Figure S4. Substrate-associated soil fungi forest diversity profiles. (A) Naïve and (B) similarity-based (phylogenetic relatedness) diversity profiles calculated from the substrate-associated soil fungi forest data.

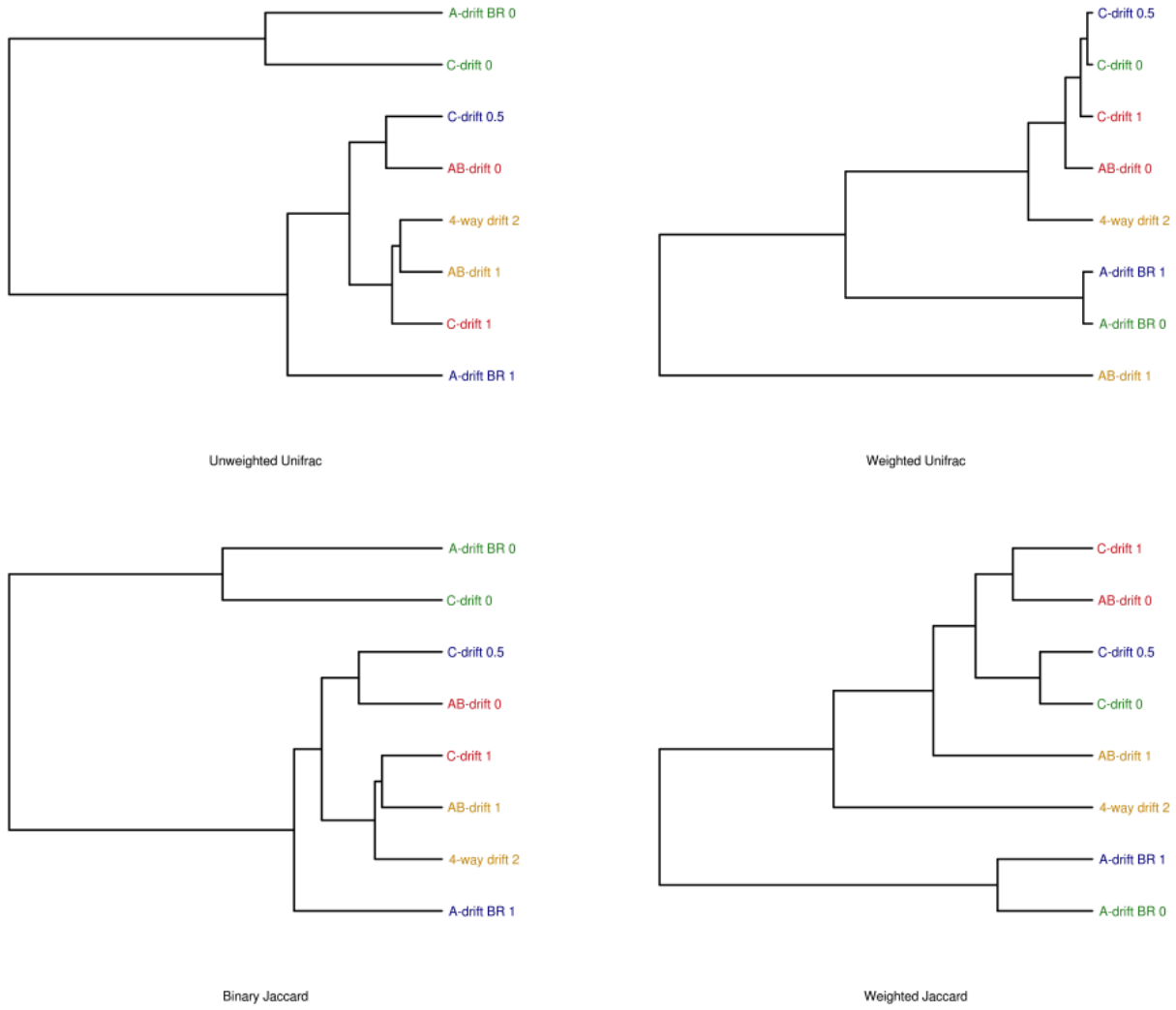


Figure S5. Acid mine drainage bacteria and archaea (HiSeq) phylogenetic (UniFrac) and taxonomic (Jaccard) hierarchical dissimilarity clusters. (Top Left) Unweighted Unifrac, (Top Right) abundance-weighted Unifrac, (Bottom Left) unweighted Jaccard, and (Bottom Right) abundance-weighted Jaccard community composition dendrograms calculated from the acid mine drainage bacteria and archaea HiSeq data.

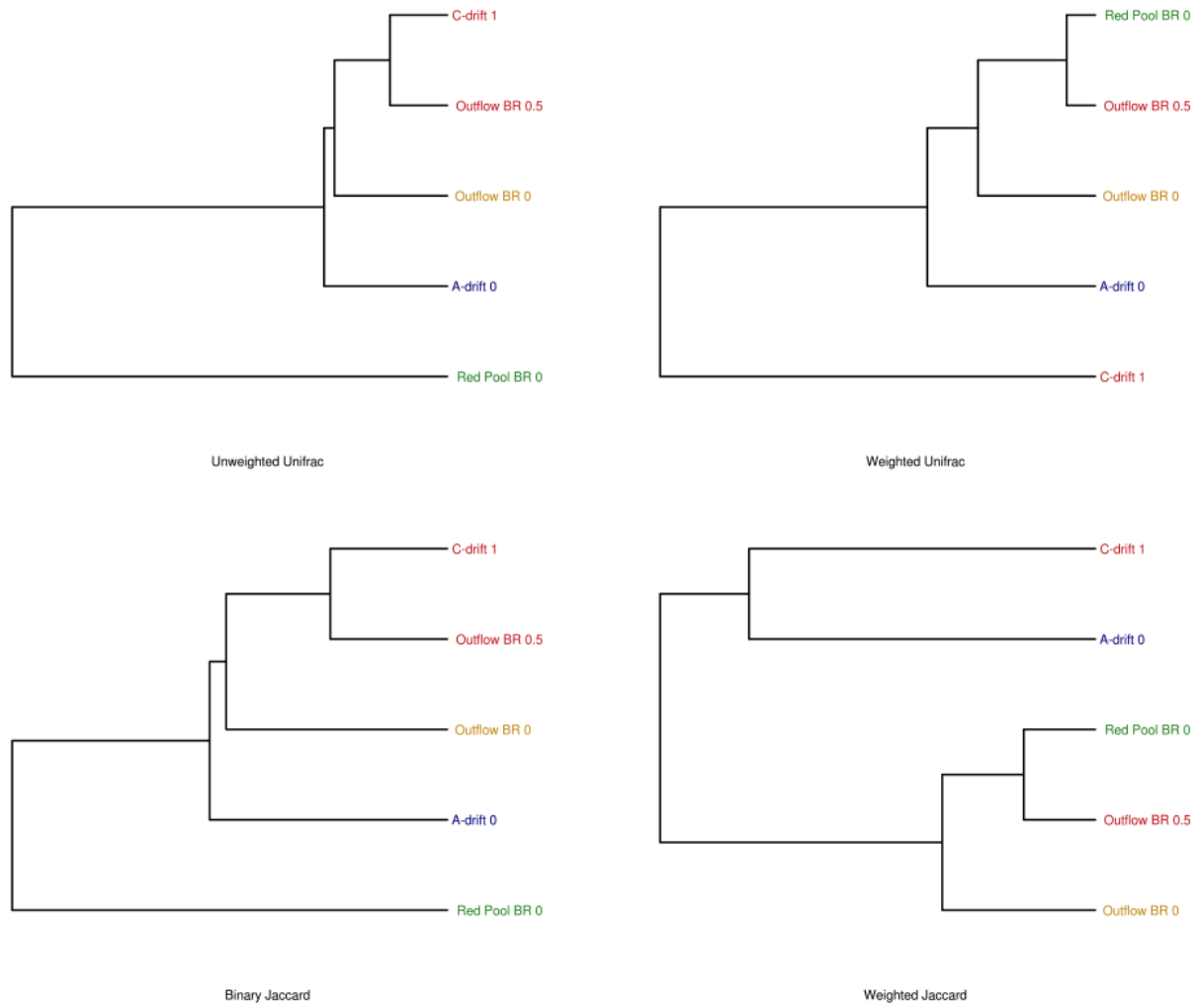


Figure S6. Acid mine drainage bacteria and archaea (GAIIx) phylogenetic (UniFrac) and taxonomic (Jaccard) hierarchical dissimilarity clusters. (Top Left) Unweighted Unifrac, (Top Right) abundance-weighted Unifrac, (Bottom Left) unweighted Jaccard, and (Bottom Right) abundance-weighted Jaccard community composition dendrograms calculated from the acid mine drainage bacteria and archaea GAIIx data.

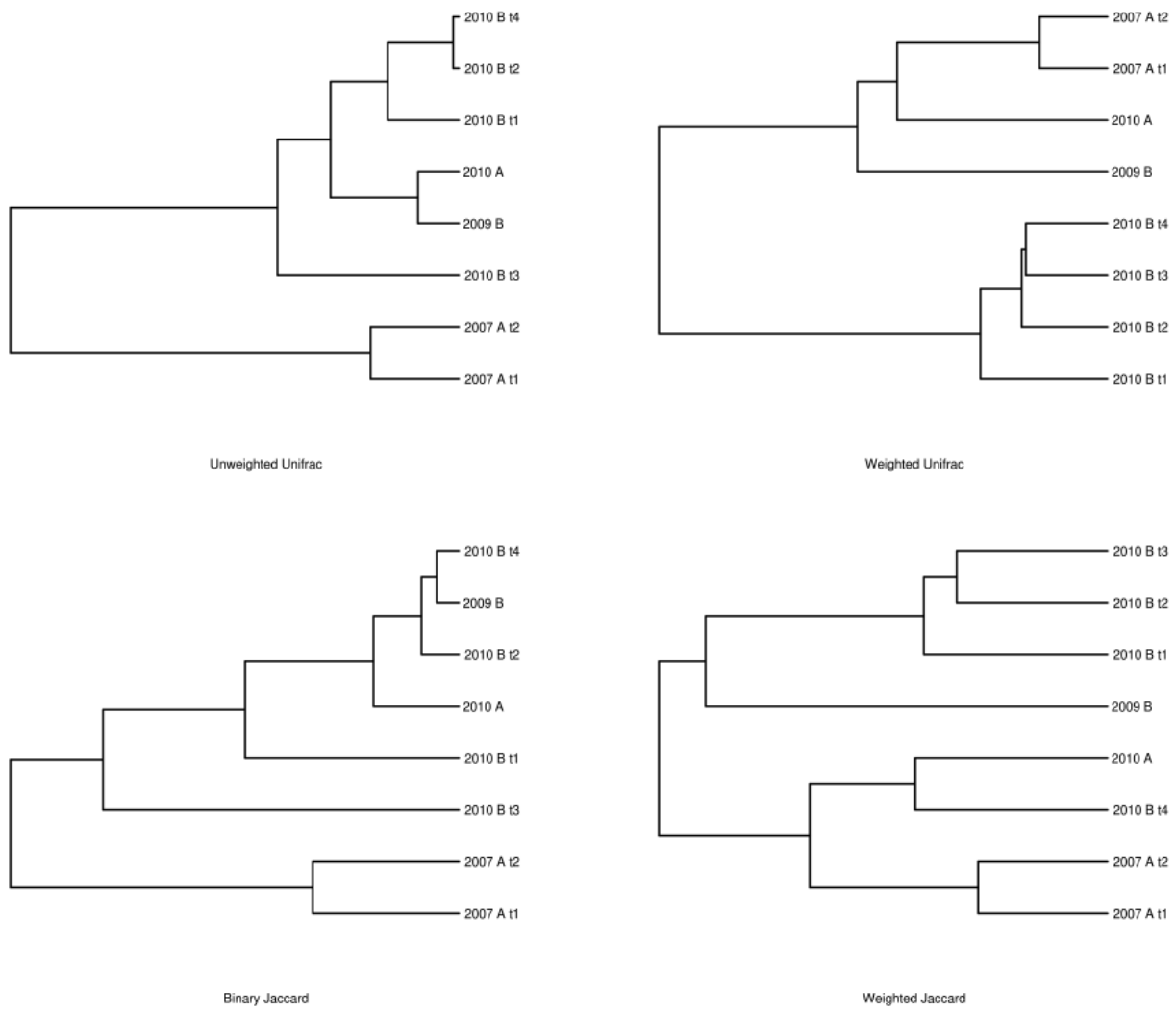


Figure S7. Hypersaline lake viruses Cluster 667 phylogenetic (UniFrac) and taxonomic (Jaccard) hierarchical dissimilarity clusters. (Top Left) Unweighted Unifrac, (Top Right) abundance-weighted Unifrac, (Bottom Left) unweighted Jaccard, and (Bottom Right) abundance-weighted Jaccard community composition dendrograms calculated from the hypersaline lake viruses Cluster 667 data.

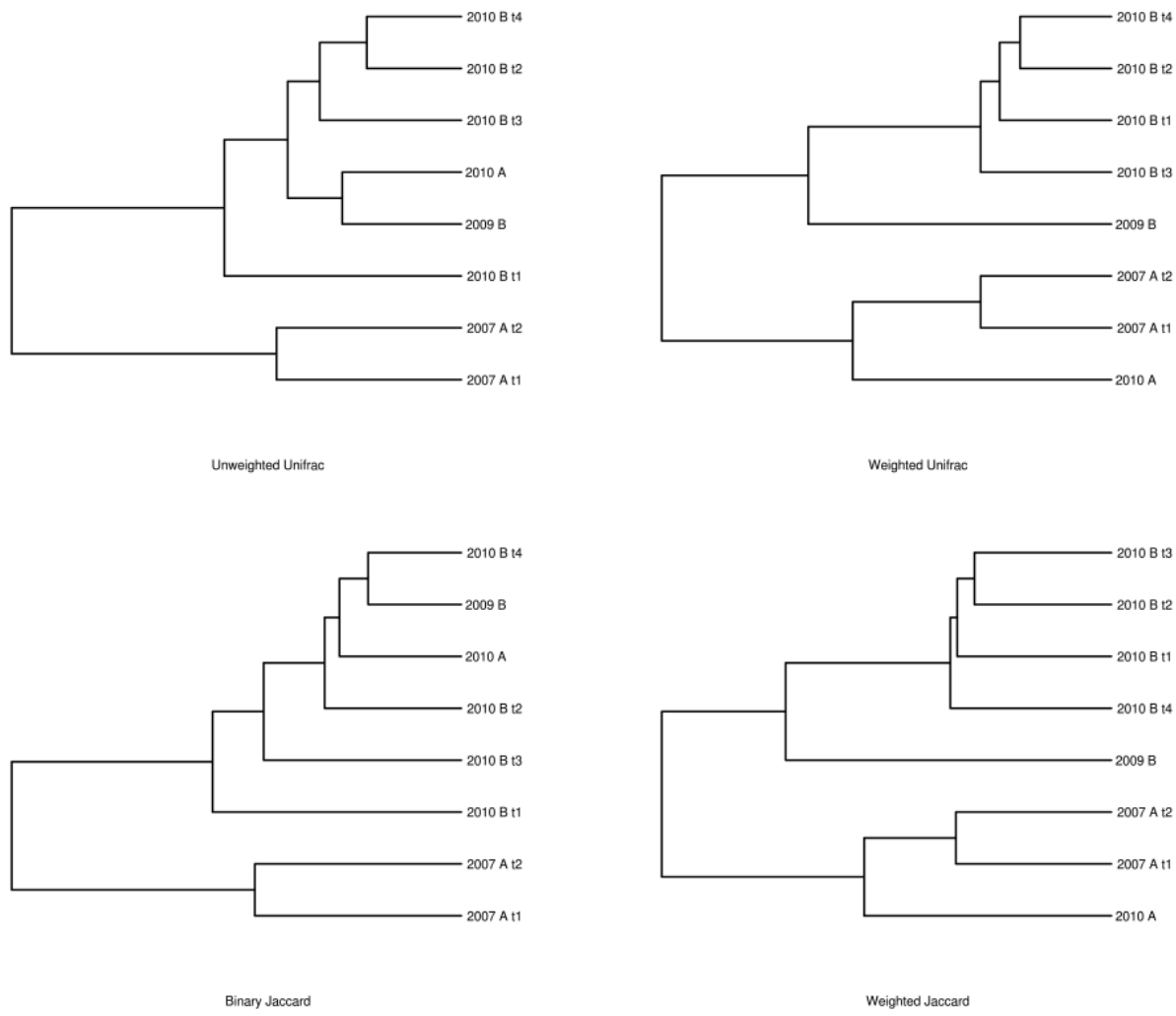


Figure S8. Hypersaline lake viruses methyltransferase phylogenetic (UniFrac) and taxonomic (Jaccard) hierarchical dissimilarity clusters. (Top Left) Unweighted Unifrac, (Top Right) abundance-weighted Unifrac, (Bottom Left) unweighted Jaccard, and (Bottom Right) abundance-weighted Jaccard community composition dendrograms calculated from the hypersaline lake viruses methyltransferase data.

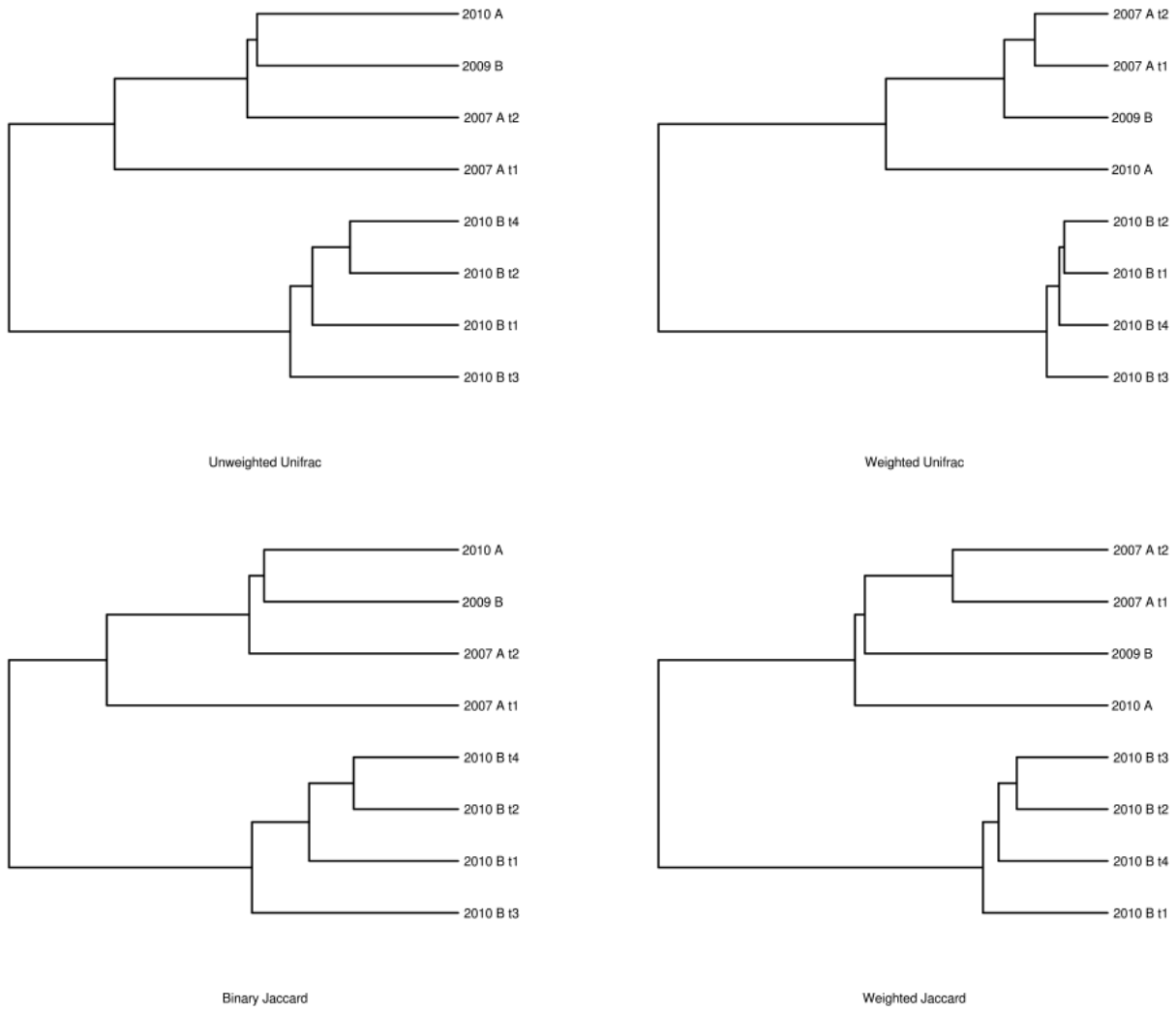


Figure S9. Hypersaline lake viruses concanavalin A-like glucanases/lectins phylogenetic (UniFrac) and taxonomic (Jaccard) hierarchical dissimilarity clusters. (Top Left) Unweighted Unifrac, (Top Right) abundance-weighted Unifrac, (Bottom Left) unweighted Jaccard, and (Bottom Right) abundance-weighted Jaccard community composition dendrograms calculated from the hypersaline lake viruses concanavalin A-like glucanases/lectins data.

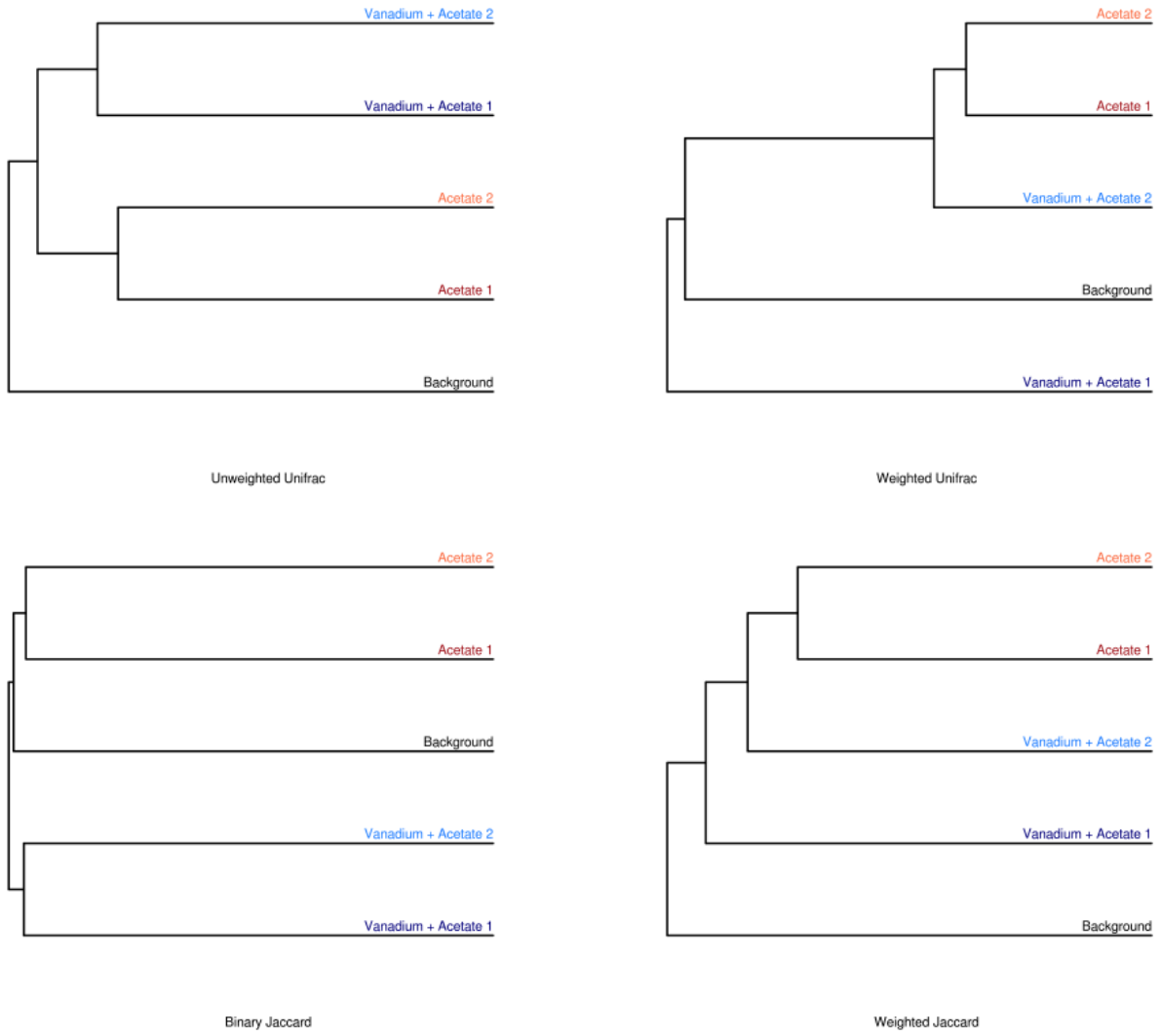


Figure S10. Subsurface bacteria phylogenetic (UniFrac) and taxonomic (Jaccard) hierarchical dissimilarity clusters. (Top Left) Unweighted Unifrac, (Top Right) abundance-weighted Unifrac, (Bottom Left) unweighted Jaccard, and (Bottom Right) abundance-weighted Jaccard community composition dendrograms calculated from the subsurface bacteria dataset.

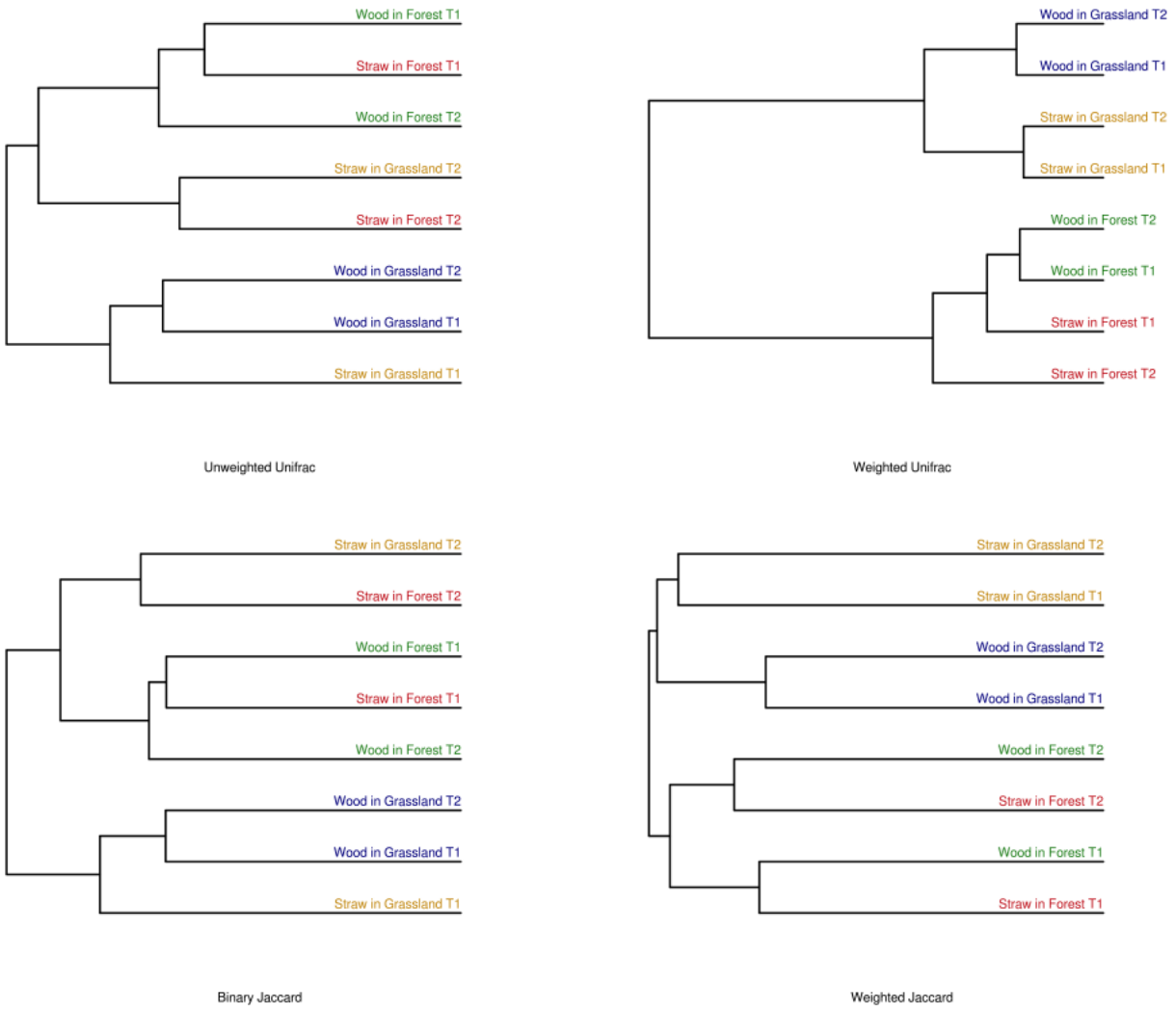


Figure S11. Substrate-associated soil fungi phylogenetic (UniFrac) and taxonomic (Jaccard) hierarchical dissimilarity clusters. (Top Left) Unweighted Unifrac, (Top Right) abundance-weighted Unifrac, (Bottom Left) unweighted Jaccard, and (Bottom Right) abundance-weighted Jaccard community composition dendrograms calculated from the substrate-associated soil fungi dataset.

CHAPTER 3. The effects of anthropogenic land-use change on multiple dimensions of soil microbial diversity in a Southeast Asian forest landscape

Abstract

Southeast Asia has the highest rate of deforestation of any major tropical region, and by 2100, the region may have lost three quarters of its primary forests. In addition to having major impacts on plant and animal diversity, tropical land-use change directly catalyzes changes to emergent ecosystem processes mediated by soil microbes. Soil bacteria, archaea, and fungi provide a number of crucial ecosystem functions in forests, including decomposition, nutrient cycling, driving plant diversity and productivity, and the mediation of cycles of the most important atmospherically reactive trace gases. This study aims to discover and quantify multiple dimensions of bacterial, archaeal, and fungal taxonomic, phylogenetic, and functional diversity in five different land-use types (primary forest, secondary forest, oil palm, rubber, and rice) throughout an dipterocarp forest landscape in Peninsular Malaysia. The objectives were to: (1) Assess the effects of anthropogenic land-use change and current land-use type on bacterial, archaeal, and fungal taxonomic and phylogenetic diversity in soils with special interest in the effects on functional genes related to CH_4, N_2O, and CO_2 cycling and to those related to phosphorus, which is often limiting in tropical soils; (2) Investigate relationships between soil microbial taxonomic diversity and local environment and spatial distance; and (3) Investigate how land-use type, soil abiotic factors, and geographic distance affect the functional gene diversity of soil microbes. The findings paint a multi-faceted picture regarding how land-use change affected soil microbial diversity, including the following key findings: the conversion of primary forest to other land-use types led to the loss of rare microbial taxa, fungal diversity was more strongly affected by land-use type than bacterial and archaeal diversity, and functional gene diversity was more strongly linked to abiotic soil environment factors than to geographic distance or land-use type. This study improves our understanding of the dimensions of microbial biodiversity by leading to greater knowledge of how anthropogenic disturbances affect soil microbial diversity.

Introduction

Tropical forest loss during the past few decades has been unprecedented, and broadly increasing, in regions throughout the world (Hansen et al. 2013). As a result, a substantial literature has developed documenting how tropical forest loss and degradation negatively impact ecosystem processes that are visible to the naked eye, such as soil erosion (Guillaume et al. 2015). Numerous studies also focus on the ways in which this land-use change affects the plant and animal diversity in these extremely biodiverse regions (e.g., Newbold et al. 2014, Laurance et al. 2012, Melo et al. 2013, Dornelas et al. 2014). However, tropical land-use change also directly catalyzes changes to emergent ecosystem processes mediated by microbes, which are not as readily visible (see Fearnside 2000).

Soil bacteria, archaea, and fungi provide a number of crucial ecosystem functions in forests, including decomposition (Setälä 2004), nutrient cycling (Arrigo 2005), driving plant diversity and productivity (Van Der Heikden et al. 2008), and the mediation of cycles of the most

important atmospherically reactive trace gases (Watling and Harper 1998). The microbial mediation of trace gases is of particular interest because wet tropical forest are the sites of the highest rates of net primary production and decomposition globally and serve as major sources of CO_2, CH_4, and N_2O (Davidson et al. 1993, Parsons et al. 1993, Silver et al. 2005, Werner et al. 2007). Thus, it is of critical importance to understand how tropical forest loss and degradation alter soil microbial communities and potentially trace gas fluxes they moderate.

In this chapter, we compare bacterial, archaeal, and fungal diversity across five land-use types (primary forest, secondary forest, oil palm, rubber, and rice) in a dipterocarp forest landscape, with special interest in the effects on functional genes related to CH_4, N_2O, and CO_2 cycling. We also focus on functional genes that are involved in the cycling of phosphorus, because it is often limiting in tropical soils (Cleveland et al. 2002), and there is the potential to see interesting and relevant changes in phosphorus cycling genes among the different land-use types. Thus, the functional genes that we specifically analyze fall into three categories: carbon, nitrogen, and phosphorus cycling and include genes specifically related to the gas flux cycles explained above, including carbon degradation, methane cycling, denitrification, nitrification, and ammonification (Table 2). This work aims to improve our understanding of the dimensions of microbial biodiversity by leading to greater knowledge of how anthropogenic disturbances affect soil microbial diversity.

Southeast Asian Land-Use Change

Southeast Asian primary rainforests are a global biodiversity hotspot (Myers 2000). While rapid deforestation and forest degradation continue in all major tropical regions of the world (Achard et al. 2002, Gibbs et al. 2010), Southeast Asia has the highest rate of deforestation of any major tropical region (Achard et al. 2002, Sodhi 2004). By 2100, Southeast Asia may have lost three quarters of its primary forests (Sodhi 2004).

Forests in Southeast Asia are primarily cleared or degraded for timber extraction, urban development or conversion to agricultural land, such as rice fields, oil palm or rubber plantations. Rice is already grown on more than 150 million hectares (Mha) worldwide (Bailey-Serres et al. 2010), and cultivation continues to increase (FAO 2011). Oil palm is the most widely grown perennial crop (currently 11 Mha), and it is projected to expand by an additional 18-26 Mha by 2050 (Corley 2009). Between 1990 and 2005, oil palm plantations in Malaysia, where our study sites were located, more than doubled to 3.6 Mha (Koh and Wilcove 2008). Rubber cultivation is also continuing to expand in Southeast Asia, with more than 500,000 ha having already been converted. It is expected that the amount of land dedicated to rubber cultivation may double or triple by 2050 (Ziegler et al. 2009).

This widespread conversion of tropical forest to agricultural land has far-reaching implications that extend beyond the local ecosystems. One major impact is the influence that land conversion has on greenhouse gas emissions. From 1850-2000, land use and land-use change caused the release of 28-40% of all anthropogenic carbon emissions (Houghton 2010). N_2O fluxes from agricultural soils account for more than 50% of all anthropogenic emissions (Robertson & Grace 2004). In our study area, oil palm plantations are regularly treated with nitrogen-based fertilizers (Corley & Tinker 2003), so while they are net CH_4 sinks (Melling et al. 2005a) and produce lower CO_2 emissions than forested sites (Melling et al. 2005b), they produce higher N_2O fluxes than forests (Hewitt et al. 2009). The fertilization of rubber plantations is also associated with N_2O emissions (Jawjit et al. 2010), and rice cultivation is the most significant soil source of CH_4 from croplands (Robertson & Grace 2004).

Microbial Determinants of CH_4, N_2O, and CO_2 Fluxes

Bacteria, archaea, and saprotrophic fungi directly mediate the production of three major trace gases of interest. CO_2 is produced in soils by a plethora of diverse heterotrophic microbes, which are phylogenetically unconstrained decomposing organisms that break down organic matter, and by autotrophic activity of living plant roots and mycorrhizae (see reviews by Raich and Schlesinger 1992, Hanson et al. 2000, Schlesinger and Andrews 2000, Kuzyakov 2006).

Soils can both produce and consume CH_4 (e.g., Topp and Pattey 1997, Segers 1998, Le Mer and Roger 2001). Net CH_4 flux is a balance between two processes: 1) methanogenesis, which is microbial production of CH_4 under anaerobic conditions, carried out only by methanogenic archaea; and 2) methanotrophy, which is microbial consumption of CH_4, carried out by methyltrophic Alpha- and Gammaproteobacteria that utilize methane as their carbon and energy sources (McDonald et al. 2005, Dutaur and Verchot 2007). Net soil CH_4 production or consumption is determined by local oxygen availability and the microbial communities found in the soil profile.

The two main pathways of N_2O production in soils are: 1) aerobic autotrophic nitrification, the stepwise oxidation of ammonia (NH_3) to nitrite (NO_2^-) and to nitrate (NO_3^-). Nitrifers are relatively phylogenetically constrained, with nitrification carried out by monophyletic groups of obligate aerobes within Beta- and Gammaproteobacteria and archaea in Crenarchaeota (Kowalchuk and Stephen 2001, Leininger et al. 2006); and 2) anaerobic denitrification, the stepwise reduction of NO_3^- to NO_2^-, nitric oxide (NO), N_2O and ultimately N_2. Denitrifers are diverse, with denitrification primarily carried out by heterotrophic bacteria, but also carried out by autotrophic bacteria, as well as archaea, and fungi (Philippot 2002; Hayatsu 2008). In this process, facultative anaerobic bacteria use NO_3^- as an electron acceptor during the respiration of organic material in the absence of oxygen (Firestone and Davidson, 1989). Finally, although microbiologists currently believe they have a relatively good understanding of the microbial bases of N_2O and CH_4 production, they have been surprised in the past by major new findings in these arenas, such as the recognition of fungal production of N_2O and the anaerobic oxidation of CH_4 (e.g., Shoun and Tanimoto 1991, Shoun et al. 1992, Aeckersberg et al. 1991).

Recent Findings

Several recent studies have begun to investigate soil microbial diversity in these rapidly changing Southeast Asian forest landscapes by comparing bacterial, archaeal, and fungal community composition in forest and agriculture sites. Despite some contradictory findings, several consistent findings emerge from multiple recent studies. For instance, multiple studies have found that primary and secondary forests (whether once- or twice-logged) have similar soil microbial community communities, implying that the soil microbial communities of secondary forests are resilient and retain most of the structure of primary forests (McGuire et al. 2015, Tripathi et al. 2014, Tripathi et al. 2016, Lee-Cruz et al. 2013). In contrast, agricultural lands, especially oil palm plantations, have been found to have significantly different soil microbial communities than either primary or secondary forests (McGuire et al. 2015, Tripathi et al. 2016, Lee-Cruz et al. 2013). Additionally, in several cases, environmental variables have been found to be stronger determinants of soil microbial community composition than land-use type (Tripathi et al. 2016, Tripathi et al. 2012, Schneider et al. 2015). This is particularly true of pH, which is well established as a strong controller of soil bacterial diversity (Lauber et al. 2009).

However, despite these recent studies, several questions remain unanswered regarding the effects of land use change on soil microbial diversity in Southeast Asian tropical forest landscapes. Many of the above studies designed their studies with no replicates of each land-use type, or they sampled in a pseudoreplicated plot design. This lack of true replicates has been previously found to affect findings in studies of tropical forest systems and thus also affect the conclusions these studies make about anthropogenic land-use change (Ramage et al. 2012). Similarly, a lack of true replicates indicates that previous studies did not compare the effects of land-use type to the effects of geographic distance on soil microbial diversity. Thus, it is possible that the effects that different land-use types had on soil microbial community composition may have been conflated with the natural variation in soil microbial communities that may be expected across geographic space (Green and Bohannan 2006). Lastly, while some previous studies in Old World forest systems have investigated how functional diversity is affected by land-use change (e.g., Zhang et al. 2014), most have focused on taxonomic diversity. There are few examples of integrated analyses of taxonomic, phylogenetic, and functional diversity in conjunction with other dimensions of microbial biodiversity in Southeast Asian forest ecosystems.

Objectives and Hypotheses

This study seeks to discover and quantify multiple dimensions of bacterial, archaeal, and fungal taxonomic, phylogenetic, and functional diversity in five different land-use types throughout a dipterocarp forest landscape in Peninsular Malaysia:

Objective 1: Assess the effects of anthropogenic land-use change on bacterial, archaeal, and fungal taxonomic and phylogenetic diversity in soils from five different land-use types (primary forest, secondary forest, oil palm, rubber, and rice).

Hypothesis 1: Bacterial, archaeal, and fungal diversity will be higher in primary and secondary forest and lower in the three agricultural land use types (oil palm, rubber, and rice).

Rationale: Differences in plant species diversity, habitat heterogeneity, and the scale, frequency, and intensity of disturbances differ among land-types. For instance, primary and secondary forests have greater habitat heterogeneity than oil palm plantations (Luskin and Potts 2011). These different characteristics typical of the land-use types will lead to differences in the types of microbes that are able to persist in each land-use type.

Objective 2: Investigate relationships between soil microbial taxonomic diversity and local environment and spatial distance.

Hypothesis 2a: Abiotic environmental factors (e.g., soil characteristics) will be more closely linked to bacterial and archaeal taxonomic diversity than to fungal diversity.

Hypothesis 2b: Geographic distance between sampling sites will have a stronger effect on the fungal taxonomic diversity than on bacterial and archaeal taxonomic diversity.

Rationale: Soil bacterial and archaeal diversity has been found to be closely linked to local environment factors, such as pH (Lauber et al. 2009, Tripathi et al. 2012). In contrast, saprotrophic fungal diversity is more strongly driven by the woody debris availability (Kruys and Jonsson 1999), which is a direct result of the aboveground plant community diversity. This plant community diversity varies greatly across geographic space in tropical ecosystems (Ricklefs 1977).

Objective 3: Investigate how land-use type, soil abiotic factors, and geographic distance affect the functional gene diversity of soil microbes, with special interest in the effects on functional genes related to CH_4, N_2O, and CO_2 cycling, as well as phosphorus.

Hypothesis 3: Functional gene diversity (with special interest in carbon, nitrogen, and phosphorus cycling genes) will be most strongly linked to abiotic environmental factors.

Rationale: Microbes that persist in a given environment must have genes that allow them to perform the functions that enable their survival. Thus, functional gene diversity will be strongly tied to the characteristics of local soil environment.

Overall, this study aims to improve our understanding of the dimensions of microbial biodiversity by leading to greater knowledge of how anthropogenic disturbances (i.e., logging, agriculture) affect soil microbial diversity. The need for such knowledge is urgent due to the rapid land use changes occurring in the region.

Methods

Sites

Soil samples were collected in five different land-use types throughout Peninsular Malaysia in June 2012. The five land-use types were primary forest, secondary forest, oil palm, rice, and rubber. Sampling in each of these land-use types was replicated in three 1 ha plots separated by at least 1 km.

Two of the primary forest plots and one of the secondary forest plots were located in the forest within and just outside of the Forest Research Institute Malaysia (FRIM) in the town of Kepong in the state of Selangor. One primary forest plot and two of the secondary forest plots were located in the Pasoh Forest Reserve (PFR) in the state of Negeri Sembilan. All three oil palm plots and all three rubber plots were located in the agricultural land surrounding the town of Mentakab in the state of Pahang. All three rice plots were located in the town of Sekinchan in the state of Selangor. See Table 1 for more details about each site.

Soil Sampling

For the primary forest, secondary forest, oil palm, and rubber plots, three soil samples (each composited from five individual cores collected down to a depth of 10 cm using a 2 cm diameter soil corer) were collected within each 1 ha plot. The locations of the cores were determined by randomly selecting a distance (either 5m, 15m, or 40m) and compass direction from the center of the plot (Table 1). A bulk soil sample was also taken from the center of each plot to measure environmental variables.

For the rice plots, one sample was taken from the edge of each 1 ha rice plot, all of which were at least 1 km away from each other. Soils were sampled from the rice fields with this different experimental design because the land managers of rice fields did not permit us to walk into their rice fields, due to the detrimental effect this would have had on their rice yields.

Soil samples were named by the first initial of the land-use type where they were collected and were numbered from 1 to 9 (e.g., the third rubber sample was named R3). In the case of rice, to differentiate from rubber samples, samples were named Ri1, Ri2, and Ri3.

Soil Analyses

Duplicate subsamples were taken from each bulk soil sample in order to measure soil pH. Field moist soil was weighed (15 g fresh weight) and made into a soil suspension with deionized water in the ratio of 1:2. The pH meter was standardized at pH 7 and 4. The remaining bulk soil samples were weighed, dried in an oven at 45°C for six days, and then re-weighed in order to calculate water content (percent water by mass). The dried bulk soil samples were then sent to the University of California, Davis Analytical Lab to measure soil texture (percentage sand, silt, and clay), total nitrogen content, total carbon content, extractable phosphorus (Bray method), and cation exchange capacity.

The soil samples used for molecular work were express shipped from Kuala Lumpur, Malaysia to Berkeley, California the same day they were collected and were kept frozen at -80°C at the University of California, Berkeley until DNA extractions were performed.

DNA was extracted from the soil samples using an adapted version of the MoBio PowerSoil DNA Isolation Kit (Carlsbad, CA). Adaptations were as follows: A second 500ul wash of Solution C5 was used to enhance the removal of PCR inhibitors. Solution C6 was heated to 55°C in order to improve DNA elution, and a second 75ul wash of Solution C6 was used in order to improve DNA recovery.

The resulting DNA was desalted (ethanol purified) to removed contaminants and quantified using NanoDrop and Qubit. The purified DNA was then sent to the Institute for Environmental Genomics at the University of Oklahoma. There, 16S and ITS Illumina MiSeq sequencing and GeoChip 5.0_60K (containing approximately 160,000 probes from 378,000 genes) analyses were performed. For the GeoChip analyses, 500 ng of DNA per sample were used for direct labeling and hybridizing to the chip.

Data Analysis

Objective 1: Singletons were removed from the 16S and ITS sequencing data, and OTUs were defined for both at a level of 97% sequence similarity. The GeoChip dataset was log transformed. GeoChip analyses were calculated for the overall GeoChip dataset as well as for subsets of the data comprised only of carbon, nitrogen, and phosphorus cycling genes (Table 2). The gene abundances of nitrogen, carbon, and phosphorus cycling genes of interest were normalized by the gene abundances in the primary forest samples (Yang et al. 2014). They were calculated as means of the nine sampling plots in each land-use type, or three plots in the case of rice. Data analyses were performed with the ape (Paradis et al. 2004), Imap (Wallace 2012), picante (Kembel et al. 2010), and vegan (Oksanen et al. 2016) packages built for R (R Core Team 2014). 16S sequences were classified by taxonomic groups using 16S Classifier (Chaudhary et al. 2015).

Cluster dendrograms based on Bray-Curtis (abundance-based data) and Jaccard (presence-absence data) dissimilarity calculations were separately calculated from both the 16S and ITS sequencing data. Detrended correspondence analysis (DCA) ordinations were plotted from the 16S, ITS, and GeoChip datasets. These dendrograms and ordinations were utilized to assess the effects of anthropogenic land-use change on bacterial, archaeal, and fungal communities in soils from the different land-use types.

Phylogenetic trees were inferred from the 16S and ITS sequence datasets using PASTA (Practical Alignment using SATé and TrAnsitivity), which produces alignments and trees for very large sequencing datasets (Mirarab et al. 2015). Faith's Phylogenetic Diversity (Faith 1992)

and OTU richness were then calculated from the 16S and ITS phylogenetic trees using the "pd" command in the Picante R package (Kembel et al. 2010, R Core Team 2014).

Diversity profiles, which have previously been shown to improve our understanding of environmental microbial diversity datasets compared to traditional diversity indices (Doll et al. 2013), were calculated for the both the 16S and ITS sequencing data using code adapted from previously published studies (Leinster and Cobbold 2012, Doll et al. 2013). Diversity profiles visualized as graphs display effective numbers of diversity, which are mathematical generalizations of previous indices that behave more intuitively and provide more meaningful percentage and ratio comparisons (Hill 1973, Doll et al. 2013). They also allow for a graphical analysis and comparison of all of multiple diversity indices simultaneously. The q parameter works as a weighted order parameter in diversity profiles so that as q increases, the weight given to rare taxa in diversity index calculations declines. For certain values of q, the diversity calculation corresponds to commonly used diversity indices. For example, $q = 0$ is equivalent to species richness, $q = 1$ is proportional to Shannon Diversity (Shannon 1948), $q = 2$ is proportional to 1/D (inverse Simpson Diversity) (Simpson 1949), and as q moves toward ∞, it is a measure of 1/Berger-Parker Evenness (Berger and Parker 1970).

Objectives 2 and 3: Canonical correspondence analysis (CCA) ordinations were run for the 16S, ITS, and GeoChip datasets, in order to explore the link between microbial community diversity and local environmental variables. The nine environmental variables included in the ordinations were soil water content (percent water by mass), pH, soil texture (percentages of sand, silt, and clay), total nitrogen content, total carbon content, extractable phosphorus (Bray method), and cation exchange capacity (Table 3, Fig. 10).

Mantel correlations were performed between a distance matrix of the location of each sampling plot and each sample's Bray-Curtis dissimilarity for the 16S, ITS, and GeoChip datasets, in order to test the affects of geographic distance on microbial community composition.

Objective 3: Functional diversity was also calculated from the GeoChip dataset in a novel way, in order to better understand how the actual diversity of individual functional gene probes varied among land-use types. For each of the selected functional genes of interest (see Table 2), the total number of probes on the GeoChip per functional gene and their relative abundance was investigated. Using the "specnumber" and "diversity" commands in the Vegan R Package (Oksanen et al. 2016, R Core Team 2014), we calculated the individual probe richness per functional gene as well as the Inverse Simpson and Shannon Diversities of the probes. For example, for the functional gene norB included on GeoChip, there were 78 individual probes that detected the presence of norB in our dataset. Therefore, we calculated the richness of norB probes as 78 and also calculated the Inverse Simpson and Shannon Diversities of norB probes using the relative abundance information obtained from GeoChip.

Results

Objective 1:

16S Taxonomic Classifications

Assigning taxonomic classifications to the 16S sequencing dataset at the levels of class revealed that the five most abundant classes throughout the dataset were Clostridia, Bacilli, Actinobacteria, Gammaproteobacteria, and Bacteroidia. The most abundant classes varied only slightly by land use type. For both primary forest and secondary forest sites, Clostridia was most important, followed by Bacilli, Gammaproteobacteria, Actinobacteria, and Bacteriodia. For oil palm, rubber, and rice, the order was almost identical, except Actinobacteria was third most abundant and Gammaproteobacteria was fourth most abundant for all three land use types.

Of the 25 classes represented in the dataset, 19 are well-represented in all of the land-use types (Fig. 1). Of interest, Deinococci were present in only rubber and rice, Coreobacteriia were present only in rice and primary forest, and Flavobacteriia were present only in primary forest. Archaea were present almost exclusively in rice and secondary forest, and Chloracidobacteria were primarily present in primary forest, oil palm, and rice, with a low relative abundance in both secondary forest and rubber.

Diversity Profiles

Diversity profiles calculated with all samples indicated that sample O1 was a strong outlier (Fig. 4, Fig. S1, Fig. S2). Therefore, the diversity profiles were recalculated with O1 removed. The recalculated 16S diversity profile (Fig. 2) shows that at $q = 0$ (the equivalent of a species richness calculation, because rare and common OTUs are weighted equally), primary forest samples were most diverse, followed by oil palm, rubber, secondary forest, and rice samples, in that order. The rice samples were much less diverse than the other four land uses. As q moves toward 1 (equivalent to the calculation of Shannon Diversity) and the profiles become less sensitive to rare taxa, the oil palm samples became the most diverse. As q moves toward ∞ (when q is between 1 and 3), the rice samples become the most diverse, followed by the oil palm, rubber, primary forest, and secondary forest samples (see Fig. 2B).

The diversity profile calculated for the ITS sequencing data (Fig. 3) shows that at $q = 0$ (a species richness calculation, because rare and common OTUs are weighted equally), primary forest samples were most diverse, followed by secondary forest, oil palm, rubber, and rice samples. As in the 16S diversity profile, the rice samples were much less diverse than the other four land uses. As q moves toward 1 (the equivalent of Shannon Diversity), the oil palm samples become more diverse than both the primary forest and the secondary forest samples. As q moves toward ∞ and the profiles become even less sensitive to rare taxa, primary forest diversity drops below that of secondary forest, rubber, and rice samples. Rubber samples also become less diverse than rice samples (see Fig. 3B).

Phylogenetic Diversity

Calculations of Faith's Phylogenetic Diversity for the 16S dataset reveal that primary forest samples were the most phylogenetically diverse, followed by oil palm, rubber, secondary forest, and rice samples, in that order. Rice samples were much less phylogenetically diverse than the other four land-use types. These calculations matched closely with the 16S OTU

richness calculations, which show that primary forest samples had the most total OTUs, followed by oil palm, rubber, secondary forest, and rice samples (Table 4).

Calculations of Faith's Phylogenetic Diversity for the ITS dataset reveal that primary forest samples were again the most phylogenetically diverse, followed by oil palm, secondary forest, rubber, and rice samples, in that order. ITS OTU richness calculations were slightly different with primary forest samples having the most total OTUs, followed by secondary forest, oil palm, rubber, and rice samples (Table 4).

Objectives 2 and 3:

Cluster Dendrograms

In the cluster dendrogram of the 16S sequencing data (Fig. 4), all of the rice samples cluster together. The majority of the primary forest and secondary forest samples cluster together, and the majority of the oil palm and rubber samples cluster together. However, some of the oil palm samples (O4, O7, O8, and O9) and some of the primary forest samples (P5, P6, P7, P8, and P9) cluster together. The rice samples cluster with a secondary forest sample (S6) as well as the aforementioned oil palm and primary forest samples. One oil palm sample (O1) is an outlier and appears as completely basal on the dendrogram. A cluster dendrogram based on Jaccard dissimilarity calculations (presence-absence data) has the same topology and reflects the same relationships among samples as the Bray-Curtis cluster dendrogram (Fig. S9). A DCA ordination created from the 16S sequencing data also shows similar clustering and relationships among the samples as the cluster dendrograms (Figs. S2, S3). A notable difference is the in the DCA ordination, the primary forest samples from the PFR sites (P1, P2, and P3) cluster together, while all of the primary forest samples from the FRIM sites (P4, P5, P6, P7, P8, and P9) cluster together).

In the cluster dendrogram for the ITS sequencing data (Fig. 5), all of the rice samples cluster together, all of the oil palm samples cluster together, and all of the rubber samples cluster together. The oil palm and rubber samples cluster with each other. All of the primary forest and secondary forest samples cluster among each other. In particular, all of the primary forest samples taken at PFR cluster with all but one of the secondary forest samples taken at PFR, while all of the primary forest samples taken at FRIM cluster with all of the secondary forest samples taken at FRIM. The rubber and oil palm cluster and the primary forest and secondary forest cluster are similar to each other, while the rice cluster is the most unique. A cluster dendrogram based on Jaccard dissimilarity calculations (presence-absence data) has the same topology and reflects the same relationships among samples as the Bray-Curtis cluster dendrogram (Fig. S10). A DCA ordination created from the ITS sequencing data also shows similar clustering and relationships among the samples as the cluster dendrograms (Fig. S4).

The cluster dendrograms for the GeoChip data are more complex. The cluster dendrogram for the entire dataset (Fig. 6), as well as the dendrograms for the carbon cycling (Fig. 7), nitrogen (Fig. 8), and phosphorus (Fig. 9) show samples from all five land use types mostly interspersed and not clustering by land use type (see also Figs. S11, S12, S13, S14). The carbon cycling cluster dendrogram (Fig. 7) and nitrogen cluster dendrogram (Fig. 8) show the strongest clustering by land use type out of the four GeoChip dendrograms, with most of the primary forest and secondary forest samples clustering together, most of the oil palm and rubber samples clustering together, and the rice samples clustering near each other. In contrast to the cluster dendrograms, the four DCA ordinations created from the full GeoChip dataset as well as

the carbon, nitrogen, and phosphorus subsets (Figs. S5, S6, S7, S8), show much tighter clustering of the samples based on land use type. In all four of these DCA ordinations, all of the rice samples cluster together near the center of the plots, the oil palm and rubber samples cluster in the bottom two quadrants, and the primary forest and secondary forest samples cluster in the left two quadrants.

Canonical Correspondence Analyses with Environmental Variables

The 16S CCA reveals that of the nine environmental variables included in the ordination, pH, percentage silt, soil water content (percent water by mass), and extractable phosphorus were the most significantly correlated with 16S diversity, in that order (Fig. 11). The ITS CCA reveals that percentage silt, soil water content (percent water by mass), extractable phosphorus, total carbon content, and total nitrogen content were most significantly correlated with ITS diversity, in that order (Fig. 12).

The CCA ordination calculated for the full GeoChip functional gene dataset (Fig. 13), as well as the CCA ordinations for the GeoChip subset datasets containing carbon cycling (Fig. S15), nitrogen (Fig. S16), and phosphorus (Fig. S17) genes, all show that of the nine environmental variables included in analysis, percent sand, pH, percent silt, and soil water content (percent water by mass) were the most significantly correlated with functional gene diversity, in that order.

Mantel Correlations

For the Mantel correlation between the geographic distance of each sampling plot and the 16S Bray-Curtis dissimilarity, the Mantel statistic r was 0.1329 (significance of 0.045). For the correlation between the geographic distance of each sampling plot and the ITS Bray-Curtis dissimilarity, the Mantel statistic r was 0.19 (significance of 0.002).

For the correlation between the geographic distance of each sampling plot and the overall GeoChip Bray-Curtis dissimilarity, the Mantel statistic r was 0.03223 (significance of 0.294). Mantel correlations were also calculated for subsets of the GeoChip dataset comprised of the genes for carbon cycling (r was 0.03498, significance of 0.287), nitrogen (r was 0.03698, significance of 0.285), and phosphorus (r was 0.04634, significance of 0.231).

Objective 3:

Functional Gene Relative Abundances

The relative abundance of nitrogen, carbon, and phosphorus cycling genes of interest varied among the five different land-use types. When normalized by gene abundances in the primary forest samples, we see several differences among land-use types (Figure 14, Figure 15, Figure 16).

For the nitrogen cycling genes, rice and secondary forest samples had elevated denitrification genes and assimilatory nitrate reduction genes (Figure 14E, Figure 14F). Secondary forest samples also had elevated amoA nitrification gene relative abundance but reduced hao nitrification gene abundance (Figure 14C). Rice and rubber samples had reduced relative abundances of nitrification genes (Figure 14C).

For the carbon cycling genes analyzed, secondary forest samples had elevated relative abundances of carbon fixation genes (aclB and CODH). Rubber samples had decreased relative abundance of the aclB carbon fixation gene (Figure 15A). Rice samples had elevated relative

abundances of methane genes (particularly pmoA) (Figure 15B), starch degradation genes (Figure 15C), and cellulose and hemicellulose degradation genes (cellobiasse and exoglucanase) (Figure 15D). Oil palm samples had elevated relative abundances of the CDH gene for cellulose and hemicellulose degradation (Figure 15D) and of the mnp lignin degradation gene (Figure 15F).

For the phosphorus cycling genes analyzed, secondary forest and rice samples had elevated phytase gene relative abundance (Figure 16). Secondary forest samples also had reduced ppk gene relative abundance (Figure 16). It is worth noting that the normalized total gene abundances from the rubber sampling plots are almost always higher than any of the land-use types, for all of the functional genes analyzed. Similarly, the oil palm normalized gene abundances are consistently below those of the primary forest samples.

Functional Gene Probe Diversity

Richness, Inverse Simpson's Diversity, and Shannon Diversity were calculated for the total number of probes for eleven functional genes of interest on GeoChip (amoA, nifh, norB, nirK, nirS, nosZ, mcrA, pmoA, CDH, endochitinase, and exochitinase) (Table 5). Rubber was calculated to be the most rich and/or even land-use type for all but one of the genes of interest. The exception was that amoA richness and evenness were highest in primary forest samples. For amoA, nifh, norB, pmoA, CDH, endochitinase, and exochitinase, richness and evenness were consistently highest in one land-use type and lowest in another. However, for nirK, nirS, nosZ, mcrA, the three different measures of richness and evenness did not agree on which land-use types were the least rich and even.

Discussion

We investigated the effects of anthropogenic land-use change by comparing bacterial, archaeal, and fungal taxonomic and functional diversity, with special interest in the diversity of functional genes related to CH_4, N_2O, and CO_2 cycling and to those related to phosphorus, which is often limiting in tropical soils (Cleveland et al. 2002) in soils from five different land-use types. Taxonomic analyses reveal that the most abundant bacterial classes were consistent across land-use types. This indicates that the most abundant classes of bacteria were consistently common in all of the soil types, despite the disparate land uses. While most of the 25 classes present in the dataset were abundant across all of the land-use types, it is worth noting that Deinococci were present in only rubber and rice. Deinococci are highly resistant to environmental hazards, such as desiccation and high temperature (Griffiths and Gupta 2007, Battistuzzi and Hedges 2009), and in the past, forests in Malaysia were often converted to agricultural land through slash-and-burn land clearing (Abdullah and Ibrahim 2002). This perhaps indicates why Deinococci were found solely in two of the disturbed agricultural land uses but not in the forest soils (Mendes et al. 2015). Additionally, Coreobacteriia and Flavobacteriia were present in only one or two land-use types. The fact that these classes were present in only one a few of the land-use types shows that despite the fact that the most abundant classes were consistently common in all of the land-use types sampled in this study, there were habitat preferences among the rarer classes of bacteria.

As hypothesized, the conversion of primary forest led to the loss of rare microbial OTUs. The 16S diversity profile shows that at the equivalent of species richness, primary forest samples

were the most diverse. As rare species are given even less weight farther down the profile, the rice samples become the most diverse, followed by the oil palm, rubber, primary forest, and secondary forest samples. The ITS diversity profile shows that at the equivalent of species richness, primary forest samples were most diverse. As rare species were given even less weight in the diversity profile, the oil palm samples become the most diverse, and primary forest samples become the least diverse fungal communities. These findings show that primary forest soils were composed of the most rare 16S and ITS OTUs, while the oil palm and rice samples were composed primarily of a large number of common OTUs but lacked the rare OTUs of the primary forest samples. Thus, this may indicate that when primary forest soils are converted to alternate land-use types, rare types of soil microbes are lost and unable to persist in the microbial community in anthropogenically altered land uses. This loss of rare bacterial, archaeal, and fungal OTUs from forest soils upon conversion of primary forest to alternate land uses has also been described in previous studies (Rodrigues et al. 2013).

The bacterial, archaeal, and fungal phylogenetic diversity of the different land-use types was similar to the taxonomic diversity, with primary forest samples having the most 16S and ITS phylogenetic diversity. This means that as primary forest is converted to alternative land-use types, phylogenetic diversity is lost among the soil microbial communities. This also implies that the rare OTUs that are lost when primary forest is converted are not just rare in terms of abundance, but they are also more phylogenetically unique than the common OTUs that remain in the soil communities after land-use conversion.

We found that fungi were more strongly affected by land-use type than bacteria and archaea were. The 16S cluster dendrograms show that the majority of the primary forest and secondary forest samples cluster together, and the majority of the oil palm and rubber samples cluster together, as would be expected based on similar land-use type, habitat characteristics, and plant community diversity (Hartmann et al. 2009, Wieland et al. 2001, Berg and Smalla 2009). However, several of the 16S samples showed inconsistent clustering with unexpected land-use types. The fungal taxonomic data shows stronger clustering that is more consistently determined by land-use type than the 16S data. In both the cluster dendrograms and DCA ordination for the ITS sequencing data, all of the rice samples cluster together, all of the oil palm samples cluster together, and all of the rubber samples cluster together. All of the primary forest and secondary forest samples cluster among each other, while the rice cluster is the most unique. The patterns of clustering of the ITS data show a significant influence of land-use type on fungal diversity. This is likely driven by the fact that the availability of woody debris that are decomposed by saprotrophic fungi vary based on the plant community of each land-use type. Bacterial and archaeal diversity have weaker connections to plant community diversity, as the availability and type of decomposing plant matter is not so critical to their metabolism.

As hypothesized, CCA ordinations and simple Mantel correlations revealed that bacterial and archaeal taxonomic diversity was driven by different environmental variables than fungal taxonomic diversity, but that bacterial, archaeal, and fungal taxonomic community diversity were all affected by geographic distance. The CCA ordinations revealed that the strongest environmental influencers of 16S diversity were pH, percentage silt, soil water content, and extractable phosphorus, in that order. The strongest environmental influencers of ITS diversity were percentage silt, soil water content, extractable phosphorus, total carbon content, and total nitrogen content, in that order. These findings do not support the hypothesis that local abiotic environmental factors would be more closely linked to bacterial and archaeal diversity than to fungal diversity. The CCA ordinations show that both bacterial and archaeal diversity as well as

fungal diversity have multiple soil characteristics that are strongly linked to them. However, the ordinations do show that different soil characteristics drive the diversity of the different soil microbial communities, as would be expected based on their disparate life cycles and environmental needs.

The most significant abiotic influence of bacterial and archaeal diversity was pH, which is well established as a strong determinant of prokaryotic soil community composition globally (Lauber et al. 2009). pH has also previously been found to drive niche portioning in soil bacterial and archaeal diversity in Malaysian soils (Tripathi et al. 2012, Tripathi et al. 2013). After pH, silt had the next most significant effect on bacterial and archaeal diversity, followed by water content and Bray phosphorus. Silt, which decreases pore connectivity, has been found to alter soil microbial communities and lead to specific microbe-particle interactions associated with silt (Carson et al. 2010, Sessitsch et al. 2001). Lower water content, which is affected by the decreased pore connectivity that silt causes, has also been found to alter the diversity of a complex bacterial community in soil (Drenovsky et al. 2004, Carson et al. 2010), while phosphorus availability is one of the limiting factors for soil microbial growth in tropical forests (Liu et al. 2012).

The strongest local environmental influencer of fungal diversity was the percentage of silt, one aspect of soil texture. Soil texture and soil water content, which is impacted by soil texture, are both known to be drivers of fungal diversity and community composition (Griffin 1963, Landis et al. 2004). Bray phosphorus, carbon content, and nitrogen content all also had a significant impact on fungal diversity. These findings correspond to studies showing that changes in phosphorus (Beauregard et al. 2010, Treseder and Allen 2002), carbon (Broeckling et al. 2007, Nielsen et al. 2010), and nitrogen (Treseder and Allen 2002, Frey et al. 2004) availability all impact soil fungal community composition.

Taxonomic diversity was structured by geographic distance. DCA ordinations and simple Mantel correlations support the hypothesis that fungal diversity would be more strongly determined by geographic distance than bacterial and archaeal diversity would be, due to the natural variation of tree species across space in tropical forest ecosystems (Ricklefs 1977), and the close reliance of fungi on plant inputs of nutrients and substrates. Previous studies have also found that fungal diversity has high spatial variability in forest ecosystems, likely due to the patchy distribution of nutrients and preferred substrates (Luis et al. 2005). A recent study in three major tropical forest types in the western Amazon found a high degree of spatial variability related to forest type and strong correlations between the alpha and beta diversity of trees and soil fungi (Peay et al. 2013). While the Mantel correlations revealed a stronger relationship between distance and fungal diversity, bacterial and archaeal diversity was also significantly correlated to spatial distance. This result is similar to other recent studies that have found that soil bacterial community similarity decayed with distance (i.e., Monroy et al. 2012). However, these findings are in contrast to a recent study in the Amazon that found that local soil bacterial diversity increases after conversion, but that communities become more similar across space (Rodrigues et al. 2013).

As hypothesized, we found that functional gene diversity was most strongly linked to abiotic environment, when also compared to land-use type and geographic distance. Geographic distance was not a significant determinant of functional diversity. Simple Mantel correlations show that geographic distance did not have a significant effect on any of the GeoChip functional gene datasets that were examined (the full GeoChip dataset, and the nitrogen, carbon, and phosphorus cycling subsets). As we hypothesized, this means that the functional genes present in

a soil sample were not more similar to samples taken at a nearby location than to other soil samples taken farther away. Microbes that persist in a given environment must have genes that allow them to perform the functions that enable their survival, but most microbes possess some functional gene redundancy (Allison and Martiny 2008). Thus, functional diversity is more likely to be driven by the functional needs of microbial communities than by geographic distance.

Zhou et al. (2008) used GeoChip to analyze soil microbial taxa area relationships. Similar to our analysis of our GeoChip data, they found that forest soil microbes had a relative flat gene-area relationship, implying that geographic distance had little influence on microbial diversity patterns. Zhou et al. hypothesized that unexplained genetic variation could be due to unmeasured biotic or abiotic environmental factors or due to the spatial scale used for sampling. They suggested that a studying the effects of geographic distance at a much larger spatial scale (such as tens of thousands of kilometers) or at a much smaller spatial scale (such as an individual meter, or even at the scale of soil particles) may better elucidate the effects of habitat heterogeneity, since natural selection on microbes most likely occurs at different scales than most studies measure. Our study also would have benefited from sampling at these disparate spatial scales in order to determine if geographic distance did have an effect on functional gene diversity, just not at the spatial scale we sampled at.

Functional gene relative abundance was inconsistently affected by land-use type. In contrast to the findings for the 16S and ITS taxonomic data, the effects of land-use type on the GeoChip functional gene data as shown by the DCA ordinations and cluster dendrograms were not as strong. The four DCA ordinations created from the full GeoChip data show some loose clustering of the samples based on land use type, with the rice samples clustering together near the center of the plots, the oil palm and rubber samples clustering in the bottom two quadrants, and the primary forest and secondary forest samples clustering in the left two quadrants. The cluster dendrograms for the entire dataset as well as for the phosphorus subset show samples from all five land-use types mostly interspersed and not clustering by land use type. However, the carbon cycling and nitrogen cycling dendrograms show most of the primary forest and secondary forest samples clustering together, most of the oil palm and rubber samples clustering together, and the rice samples clustering near each other.

This implies that the effects of land-use type were less strong on functional gene diversity than on taxonomic diversity. We see that functional genes were loosely influenced by land-use type and the functional genes found in one sample were sometimes more similar to those found in another sample of the same land-use type. However, we fail to see the consistent clustering that we see for the 16S and ITS datasets, where taxonomic diversity in a soil sample was much more similar to that of samples of the land-use type than to samples of a different land-use type. In contrast, Paula et al. (2014) utilized GeoChip 4.0 to analyze how the conversion of Amazon rainforest to pasture affected soil microbial functional diversity. They found that pasture soils had significantly lower functional gene richness and diversity than primary forest soils. Primary forest soils and secondary forest soils also have differences in gene composition.

Similar to the findings of the DCA ordinations and cluster dendrograms, direct comparisons of the relative abundance of functional genes of interest reveal that land-use type inconsistently affects functional diversity. For the nitrogen cycling genes, rice and secondary forest sample had elevated denitrification genes and assimilatory nitrate reduction genes. Secondary forest samples also had elevated amoA nitrification gene relative abundance but reduced hao nitrification gene abundance. Rice and rubber samples had reduced relative abundances of nitrification genes. Zhang et al. (2014) used GeoChip 4.0 to compare functional

gene abundances between mature forest and secondary forest and found that land-use type had a significant effect on the signal intensities of genes related to nitrogen cycling. However, they found that the abundance of several nitrogen cycling genes, including those related to nitrogen fixation, nitrification, denitrification, dissimilatory N reduction, ammonification, and assmilatory N reduction, were all consistently higher in secondary forest soil than in primary forest soil. We found that some nitrogen cycling genes were both more and less abundant in the other land-use types when compared to primary forest soils.

For the carbon cycling genes analyzed, secondary forest samples had elevated relative abundances of carbon fixation genes (aclB and CODH), indicating carbon fixation mediated by soil microbes was higher in the secondary forest soils compared to the primary forest soils. Similarly, Zhang et al. (2014) found that functional genes related to carbon fixation had higher signal intensity in secondary forest compared to primary forest. In contrast, rubber samples had decreased relative abundance of the aclB carbon fixation gene, indicating reduced soil microbial carbon fixation compared to the primary forest soils. Rice samples had elevated relative abundances of methane genes (particularly pmoA), as would be expected based on the high soil water content of the rice soil samples. Compared to the primary forest soils, rice samples had elevated starch degradation genes, and cellulose and hemicellulose degradation genes (cellobiasse and exoglucanase), and oil palm samples had elevated relative abundances of the CDH gene for cellulose and hemicellulose degradation and of the mnp lignin degradation gene. Zhang et al. (2014) also found that the soil microbial functional genes related to carbon degradation were significantly different between forest types, with their relative abundances being higher in secondary forest than in primary forest. For the phosphorus genes analyzed, secondary forest and rice samples had elevated phytase gene relative abundance compared to the primary forest samples. Secondary forest samples also had reduced ppk gene relative abundance, both of which indicate land-use type affected soil microbial metabolic activity related to phosphorus cycling.

Calculations of richness, Inverse Simpson's Diversity, and Shannon Diversity for eleven functional genes of interest did not reveal any clear trends regarding the relationship between land-use type and the richness and evenness of the GeoChip probes related to selected carbon and nitrogen genes of interest. For instance, mcrA and pmoA genes are mediate methane cycling, and we thus hypothesized that the richness of these gene probes would be higher in the rice samples, which had much higher moisture content than soils from the other land-use types. However, richness, Inverse Simpson's Diversity, and Shannon Diversity of the mcrA probes were actually lowest in the rice samples. The richness of pmoA probes was also lowest in rice, and the Inverse Simpson's Diversity and Shannon Diversity of pmoA probes were close to lowest in the rice samples. Additionally, contrary to predictions, rubber was calculated to be the most rich and/or even land-use type for all but one of the genes of interest. This may be an artifact of the GeoChip methods of measuring probe abundance, rather than have a clear biological explanation. This seems to be the case because both the functional gene probe diversity and the normalized total gene abundances from the rubber sampling plots were almost always higher than any of the land-use types. Similarly, the oil palm normalized gene abundances were consistently lower than the other land-use types. These trends likely imply that the background signal intensities in the GeoChip data were overall higher for rubber and lower for oil palm than for the land-use types. This impedes our ability to make a direct comparison of the probe diversity and relative gene abundances of the five land-use types.

Functional gene diversity was most strongly linked to abiotic environment. The CCA ordinations of functional gene show diversity was most strongly linked to pH, to two measures of soil texture (percentage sand and percentage silt), and to soil water content. This was true for the entire GeoChip dataset, as well as for the carbon, nitrogen, and phosphorus cycling subsets. Like taxonomic diversity, pH has been previously found to drive functional gene diversity in soil microbes (Zhang et al. 2013). Soil texture and water content have also been previously shown to impact functional gene diversity. Reeve et al. (2010) studied soil microbial gene frequency and diversity in an agro-ecosystem and found that denitrification potential was greater in organically managed fine-textured soils, but not in coarse-textured soils. Drenovsky et al (2004) found soil water content was strongly linked to the microbial community composition of lowland rice soils. Soil water content influences microbial diversity by impacting oxygen availability and nutrient availability. In general, high water content reduces soil oxygen availability, leading to an increase in anaerobic microbes with functional genes allowing their survival in low oxygen conditions, while low water content lowers microbial activity (Drenovsky et al. 2004, Carson et al. 2010).

Surprisingly, total carbon, total nitrogen, and Bray phosphorus, which were found to vary among the land-use types, were not among the top abiotic factors influencing functional gene diversity for the carbon, nitrogen, and phosphorus cycling subsets, respectively. This could be due to relatively similar levels of total carbon, total nitrogen, and Bray phosphorus among the different soil samples. Without large variation in their availability among the different soil types, they were not seen as drivers of differences in functional gene diversity.

Future work could better tease apart the relationship between soil microbial diversity in this ecosystem and geographic distance by sampling at a greater variety of spatial scales. Further phylogenetic methods of diversity assessment, such as spatial phylogenetics (Mishler et al. 2014, Thornhill et al. 2016), could also be added. Additionally, it would be valuable to conduct a longitudinal study in which soil from the same area is sampled before, during, and after a primary forest is converted to an alternate land-use type. This would reduce the effects that the naturally high variability of plant communities across spcace in tropical forest systems has on the soil microbial communities sampled and allow us to better isolate the effects of land-use change. This study, and any future continuations of this work, can lead to a greater understanding of how anthropogenic disturbances affect soil microbial diversity. The need for such knowledge is urgent due to the rapid land-use changes occurring in all major tropical forest ecosystems.

Table 1. Detailed information about sampling sites

Sample	Land Use Type	Plot Location	City/Town	State	Sampling Distance from Plot Center (m)	Sampling Direction from Plot Center	Litter Type	Notes
P1	Primary Forest	PFR	N/A	Negeri Sembilan	5	South	Leaf litter	
P2	Primary Forest	PFR	N/A	Negeri Sembilan	15	East	Leaf litter	Compartment 22
P3	Primary Forest	PFR	N/A	Negeri Sembilan	40	West	Leaf litter	
P4	Primary Forest	FRIM	Kepong	Selangor	5	West	Leaf litter	Sampling was beyond the canopy walkway (outside FRIM)
P5	Primary Forest	FRIM	Kepong	Selangor	15	North	Leaf litter	
P6	Primary Forest	FRIM	Kepong	Selangor	40	East	Leaf litter	
P7	Primary Forest	FRIM	Kepong	Selangor	5	South	Leaf litter	Sampling was beyond the canopy walkway (outside FRIM)
P8	Primary Forest	FRIM	Kepong	Selangor	15	West	Leaf litter	
P9	Primary Forest	FRIM	Kepong	Selangor	40	North	Leaf litter	
S1	Secondary Forest	PFR	N/A	Negeri Sembilan	5	West	Leaf litter	Compartment 21 (logged in 1955)
S2	Secondary Forest	PFR	N/A	Negeri Sembilan	15	South	Leaf litter	
S3	Secondary Forest	PFR	N/A	Negeri Sembilan	40	North	Leaf litter	
S4	Seconday Forest	PFR	N/A	Negeri Sembilan	5	North	Leaf litter	Compartment 21 (logged in 1955)
S5	Secondary Forest	PFR	N/A	Negeri Sembilan	15	West	Leaf litter	
S6	Secondary Forest	PFR	N/A	Negeri Sembilan	40	East	Leaf litter	Wetter soil than other plots
S7	Secondary Forest	FRIM	Kepong	Selangor	5	West	Leaf litter	Sampling plots were located off of the Keruing Trail
S8	Secondary Forest	FRIM	Kepong	Selangor	15	East	Leaf litter	
S9	Secondary Forest	FRIM	Kepong	Selangor	40	South	Leaf litter	
O1	Oil Palm	Agriculture	Mentakab	Pahang	5	East	Some moss	Smallholder plot, 22-24 years old, recent herbicide application, rocky red soil, planted on hill
O2	Oil Palm	Agriculture	Mentakab	Pahang	15	North	Leaf litter	
O3	Oil Palm	Agriculture	Mentakab	Pahang	40	South	Moss	

O4	Oil Palm	Agriculture	Mentakab	Pahang	5	East	Bare soil	Smallholder plot, 20 years old, minor herbicide use, sandy slightly rocky yellow soil, slight hill
O5	Oil Palm	Agriculture	Mentakab	Pahang	15	North	Rocks & moss	
O6	Oil Palm	Agriculture	Mentakab	Pahang	40	South	Moss	
O7	Oil Palm	Agriculture	Mentakab	Pahang	5	West	Understory vegetation	Large plantation, 20 years old, no herbicide use, very wet soil
O8	Oil Palm	Agriculture	Mentakab	Pahang	15	South	Understory vegetation	
O9	Oil Palm	Agriculture	Mentakab	Pahang	40	North	Understory vegetation	
R1	Rubber	Agriculture	Mentakab	Pahang	5	South	Leaf litter	Smallholder plot, 4 years old, recent herbicide application, rocky red soil, planted on hill
R2	Rubber	Agriculture	Mentakab	Pahang	15	East	Leaf litter	
R3	Rubber	Agriculture	Mentakab	Pahang	40	West	Leaf litter	
R4	Rubber	Agriculture	Mentakab	Pahang	5	East	Leaf litter	Smallholder plot, 7 years old, minor herbicide use, sandy yellow soil with few rocks, flat, originally planted in 1907
R5	Rubber	Agriculture	Mentakab	Pahang	15	North	Leaf litter	
R6	Rubber	Agriculture	Mentakab	Pahang	40	South	Leaf litter	
R7	Rubber	Agriculture	Mentakab	Pahang	5	West	Leaf litter	Large plantation, 13-16 years old, recent herbicide application, root rot disease present in some trees
R8	Rubber	Agriculture	Mentakab	Pahang	15	South	Leaf litter	
R9	Rubber	Agriculture	Mentakab	Pahang	40	North	Leaf litter	
Ri1	Rice	Agriculture	Sekinchan	Selangor	Edge	Edge	Recently harvested rice	Large rice plantations
Ri2	Rice	Agriculture	Sekinchan	Selangor	Edge	Edge	Recently harvested rice	
Ri3	Rice	Agriculture	Sekinchan	Selangor	Edge	Edge	Recently harvested rice	

Table 2. GeoChip gene category subsets and the subcategories they contain

GeoChip Gene Category Subsets	Subcategories Included
Carbon	Carbon degradation
	Carbon fixation
	Methane
Nitrogen	Ammonification
	Assimilatory nitrate reduction
	Denitrification
	Dissimilatory nitrate reduction
	Nitrogen assimilation
	Nitrification
	Nitrogen fixation
Phosphorus	Phytic acid hyrolysis
	Polyphosphate degradation
	Polyphosphate synthesis

Table 3. Soil environment variables, as measured from bulk soil samples taken from the center of each sampling plot

Site	TotalN	TotalC	BrayP	CEC	%Sand	%Silt	%Clay	pH	Water Content (% H_2O by mass)
P1	0.058	0.77	2.1	3.1	76	15	9	4.63	5.93
P2	0.079	1.08	3.2	6.1	75	13	12	4.27	7.46
P3	0.09	1.21	2.7	7.7	59	26	15	4.49	11.69
P4	0.154	1.6	4.8	7.2	68	15	17	5.15	13.95
P5	0.193	1.91	7.2	9.6	71	16	13	5.56	14.06
P6	0.175	2.02	20	8.5	79	9	12	5.63	12.75
P7	0.193	2.07	13.5	10.1	72	16	12	5.58	12.29
P8	0.228	2.27	21.8	10.9	72	15	13	5.46	15.38
P9	0.116	1.22	1.8	7.8	73	15	12	5.96	14.21
S1	0.212	2.57	1.8	11.6	32	42	26	4.2	16.92
S2	0.151	1.9	0.9	9.9	36	43	21	4	14.88
S3	0.28	3.35	1.6	12.9	24	44	32	3.91	19.93
S4	0.178	2.47	1.9	12.4	15	57	28	3.88	17.11
S5	0.218	3.03	1.9	11.2	37	49	14	4.31	16.97
S6	0.131	1.98	1.4	10.5	41	44	15	4.76	30.35
S7	0.162	2.04	5.1	11.3	58	8	34	3.78	27.90
S8	0.151	1.87	4.8	10.9	63	8	29	3.95	24.48
S9	0.14	2.03	2.9	11.1	66	4	30	3.65	23.01
O1	0.253	2.34	330	17	38	24	38	4.2	16.53
O2	0.213	2.11	9.3	15.2	25	27	48	4.02	19.00
O3	0.1	1.02	1.6	8	47	29	24	4.53	14.03
O4	0.1	0.92	41	9.2	33	29	38	5	15.06
O5	0.08	0.98	7.1	8.5	35	31	34	4.37	13.08
O6	0.091	1.02	6.4	9.4	32	32	36	4.2	12.69
O7	0.157	1.31	1	13.4	12	48	40	5.27	25.88
O8	0.242	2.35	1.8	17.8	13	45	42	5.11	34.84
O9	0.199	1.96	4.8	15	18	44	38	5.16	27.86
R1	0.164	2.13	2.9	10.3	45	33	22	4.88	17.02
R2	0.136	1.43	2.5	11	36	28	36	4.63	17.79
R3	0.124	1.22	1.4	13.3	40	26	34	4.52	20.35
R4	0.103	1.11	9.9	9	32	39	29	4.11	15.95
R5	0.131	1.4	5.8	9.9	42	34	24	4.79	19.84
R6	0.176	1.69	4.2	12.2	28	31	41	4.66	20.45
R7	0.07	0.84	15	6.5	61	30	9	4.66	14.92
R8	0.122	1.42	33.7	6.8	51	33	16	4.4	19.75
R9	0.062	0.83	50.6	5.6	57	31	12	4.81	17.85
Ri1	0.272	3.37	2	23.4	20	42	38	4.23	28.63
Ri2	0.256	5.26	47	29.5	16	36	48	5.23	34.32
Ri3	0.715	12.35	14.7	31.6	28	40	32	4.6	42.34

Table 4. Faith's Phylogenetic Diversity (PD) calculations and total number of OTUs for the 16S and ITS datasets

	16S PD	16S OTU Richness	ITS PD	ITS OTU Richness
Primary Forest	2384.796	34067	2161.5581	5868
Secondary Forest	2099.3	29474	1826.2358	4824
Oil Palm	2312.9	31611	1909.8546	4605
Rubber	2170.083	30783	1627.564	4008
Rice	1536.515	18436	895.3362	1744

Table 5. Functional Gene Probe Diversity for Selected Genes

	Primary Forest	Secondary Forest	Oil Palm	Rubber	Rice
amoA Richness	26	25	20	25	24
amoA InvSimpson	20.88603	22.04264	18.07426	20.83916	20.57163
amoA Shannon	3.113071	3.146106	2.936101	3.088165	3.081169
nifh Richness	630	629	614	647	574
nifh InvSimpson	528.9684	531.1643	516.1086	557.7379	526.7013
nifh Shannon	6.324034	6.328596	6.302786	6.370442	6.299768
norB Richness	78	78	79	81	78
norB InvSimpson	71.04056	71.11267	71.84619	75.42078	72.65592
norB Shannon	4.295299	4.295774	4.309443	4.348667	4.312588
nirK Richness	314	316	319	329	294
nirK InvSimpson	277.5359	277.432	277.2052	296.1913	277.4357
nirK Shannon	5.667888	5.668473	5.669682	5.726087	5.648915
nirS Richness	356	350	366	374	339
nirS InvSimpson	311.4359	309.8141	312.0493	332.2665	313.3288
nirS Shannon	5.785938	5.777045	5.794336	5.844095	5.777599
nosZ Richness	447	440	453	454	423
nosZ InvSimpson	395.7002	392.8181	396.4925	414.5709	395.9321
nosZ Shannon	6.020575	6.012264	6.026049	6.056648	6.006978
mcrA Richness	127	122	132	129	115
mcrA InvSimpson	109.2112	107.3364	107.9881	112.508	106.6921
mcrA Shannon	4.744451	4.721134	4.744103	4.764005	4.700374
pmoA Richness	72	76	76	87	71
pmoA InvSimpson	60.87326	63.71251	61.01901	70.94554	64.3807
pmoA Shannon	4.159475	4.209761	4.175859	4.323533	4.20226
CDH Richness	60	58	60	61	56
CDH InvSimpson	52.86736	52.52397	54.6098	56.46036	53.40248
CDH Shannon	4.004744	3.993421	4.033947	4.062115	3.998243
endochitinase Richness	230	217	231	239	213
endochitinase InvSimpson	192.9403	190.1811	194.6166	205.2268	194.4867
endochitinase Shannon	5.319094	5.295066	5.329702	5.372548	5.306541
exochitinase Richness	25	27	25	27	23
exochitinase InvSimpson	22.05088	22.22095	21.89582	23.36971	21.10473
exochitinase Shannon	3.140688	3.168689	3.143218	3.200054	3.082667

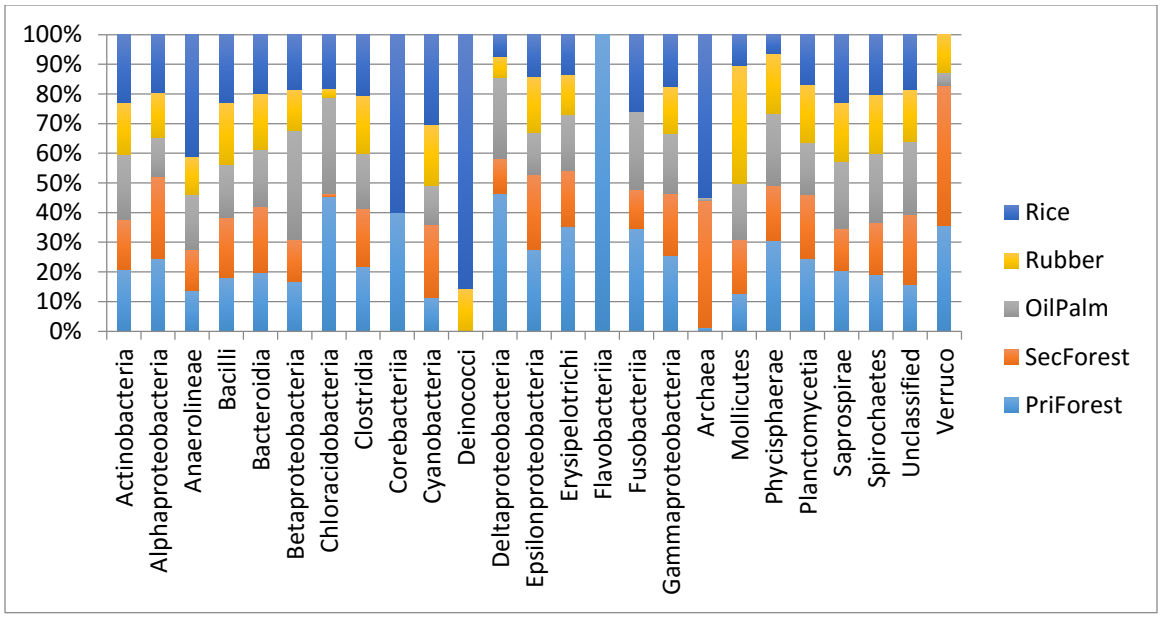
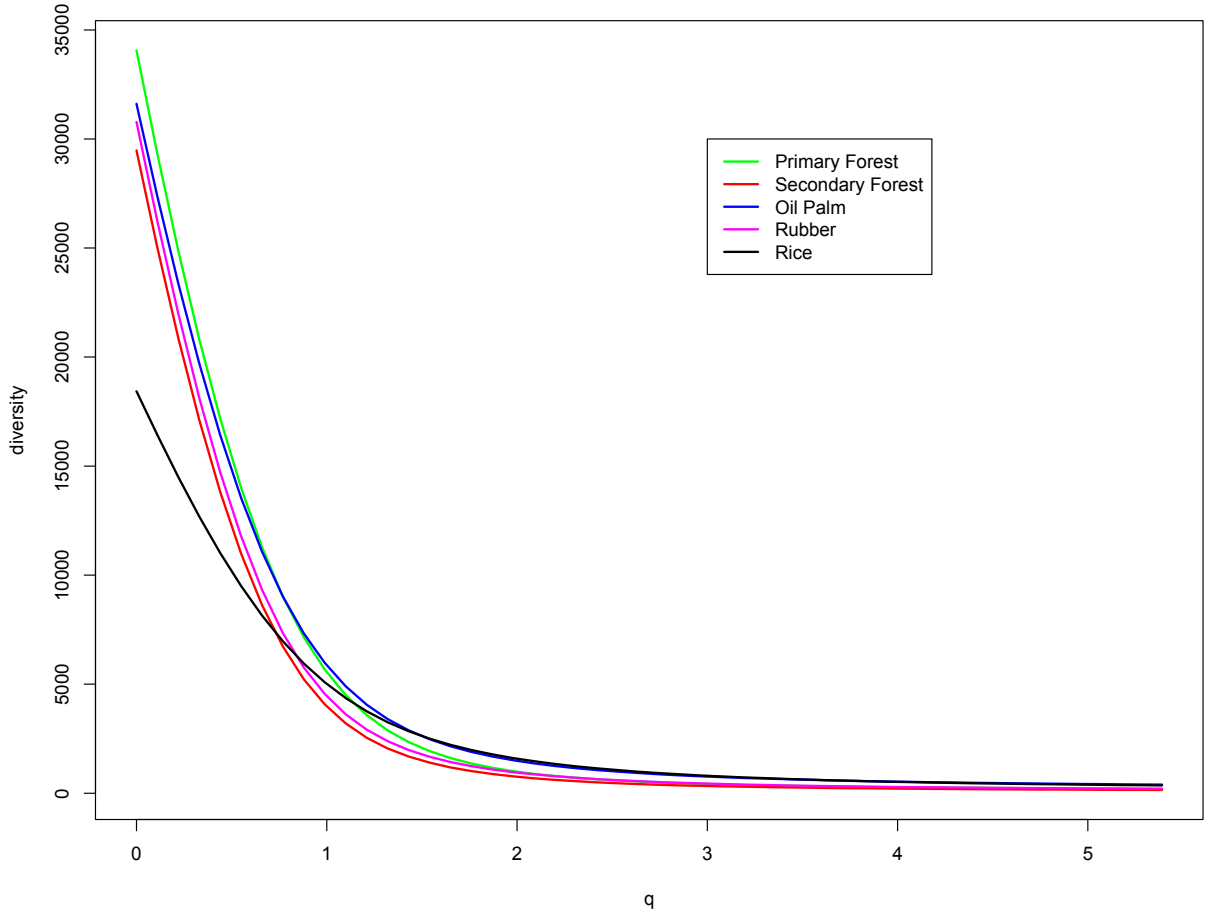


Figure 1. 16S taxonomic groups (assigned to class) by land-use type.

(A)



(B)

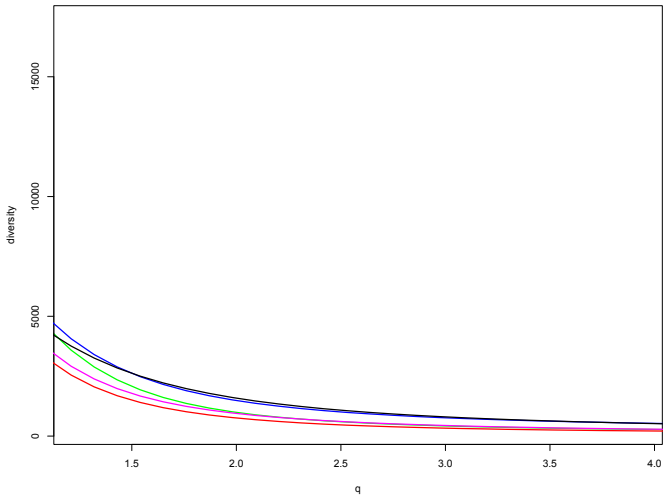
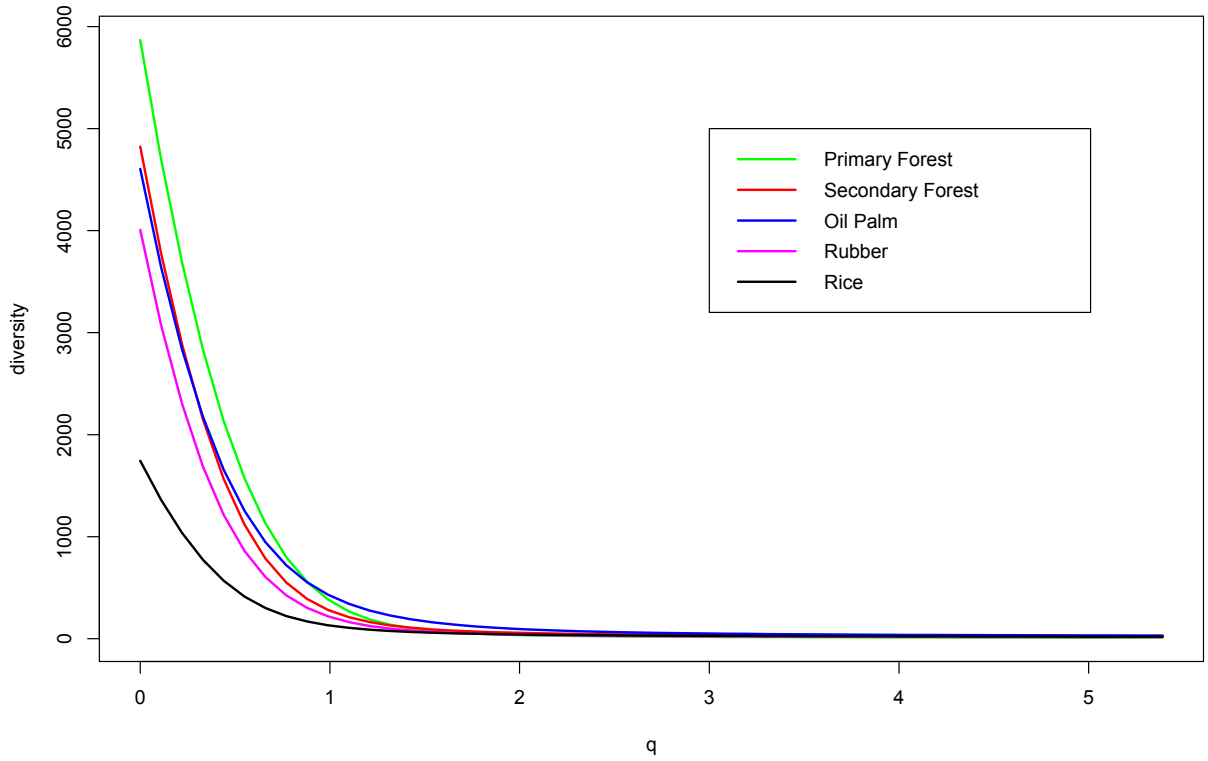


Figure 2. (A) Diversity profile calculated for the 16S dataset, excluding the outlier sample, O1. (B) A closer view of the portion of the diversity profile where $1.5 \leq q \leq 4.0$.

(A)



(B)

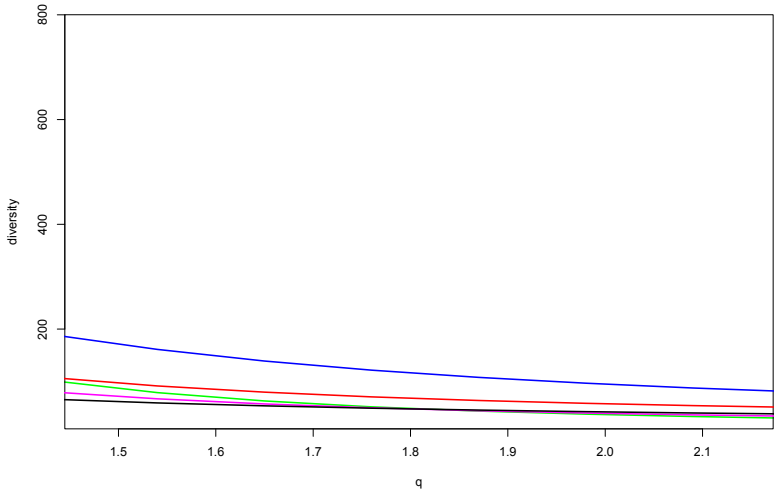


Figure 3. (A) Diversity profile calculated for the ITS dataset. (B) A closer view of the portion of the diversity profile where $1.5 \leq q \leq 2.1$.

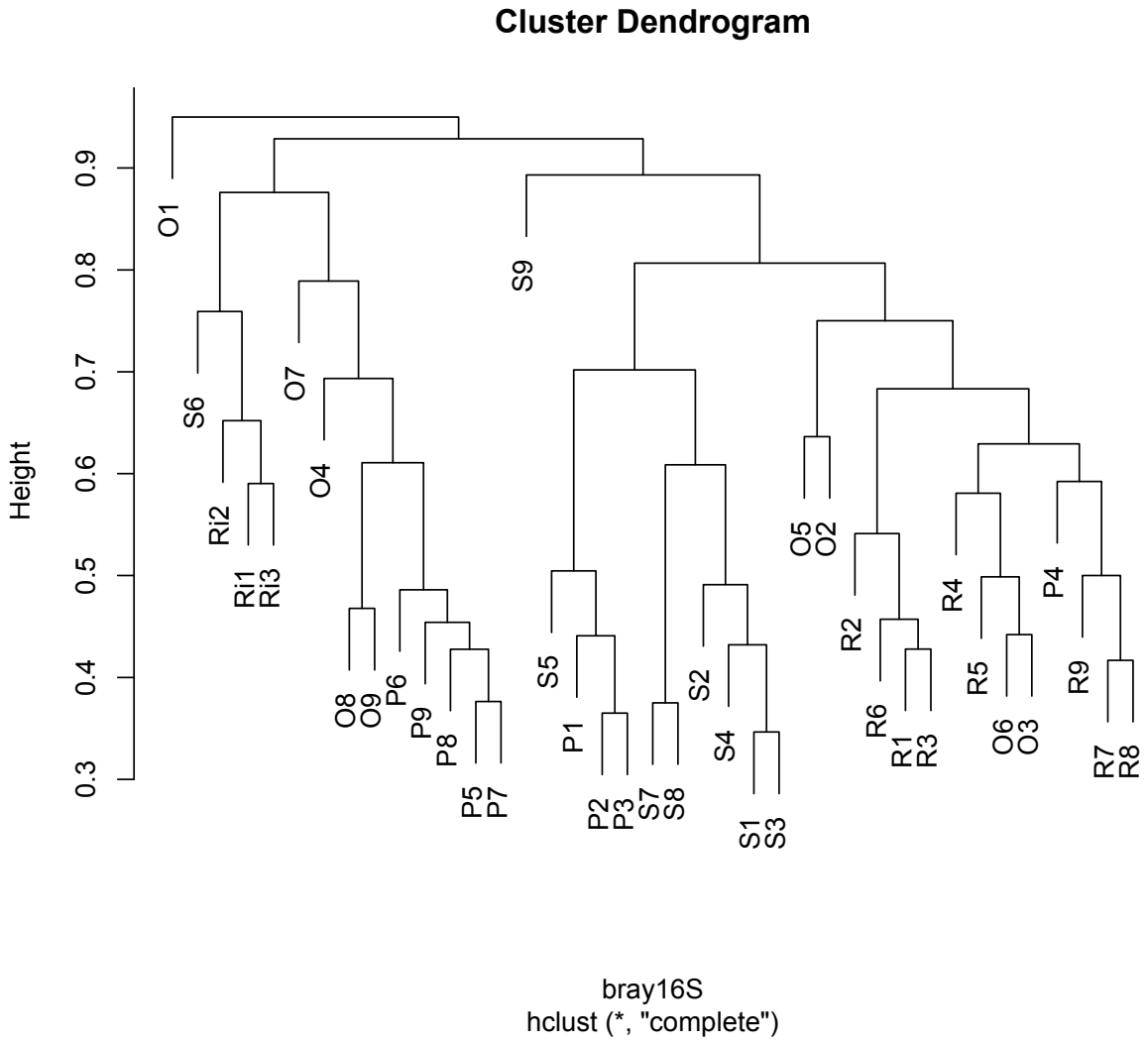


Figure 4. Cluster dendrogram of Bray-Curtis dissimilarity of the 16S dataset.

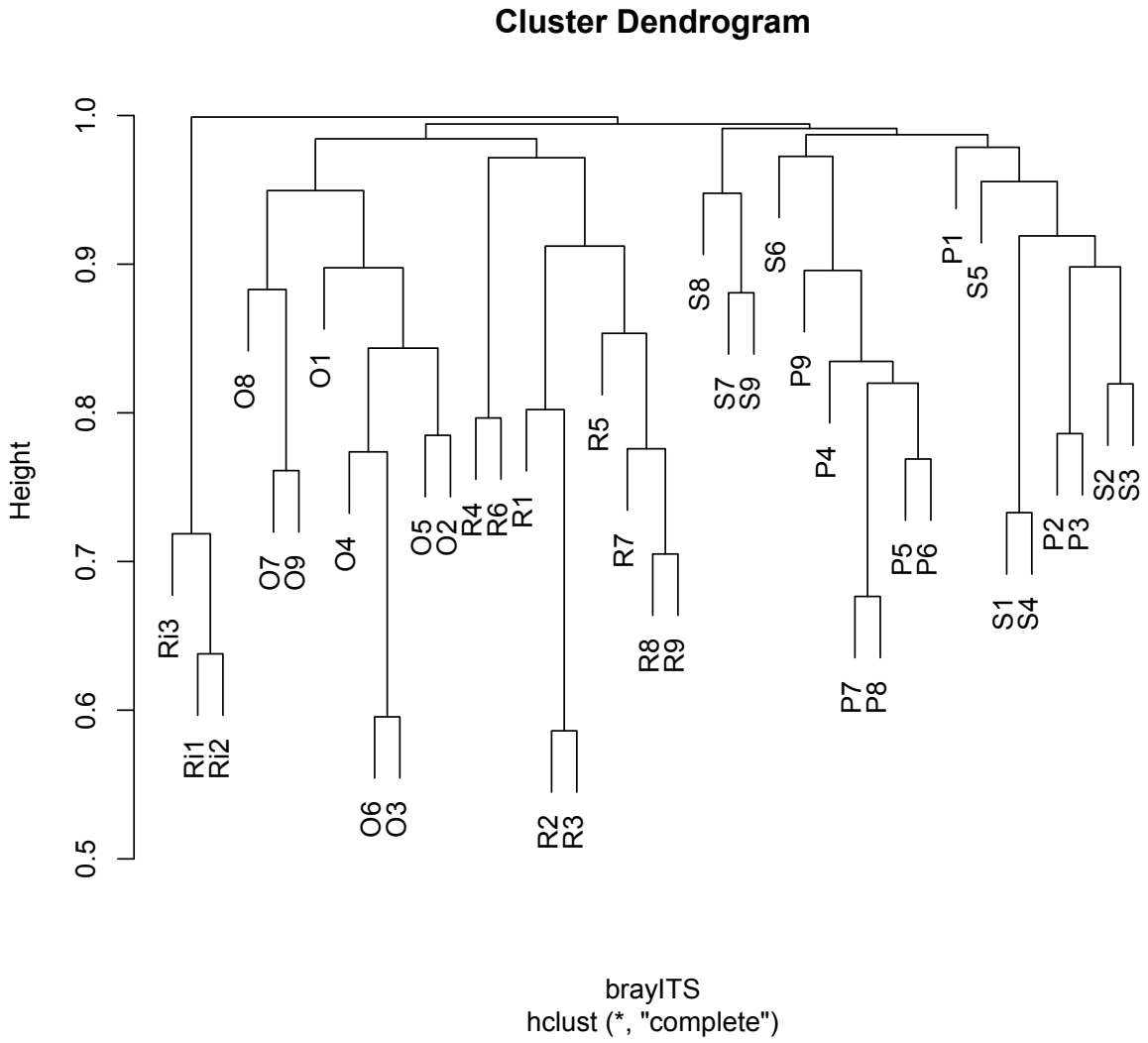


Figure 5. Cluster dendrogram of Bray-Curtis dissimilarity of the ITS dataset.

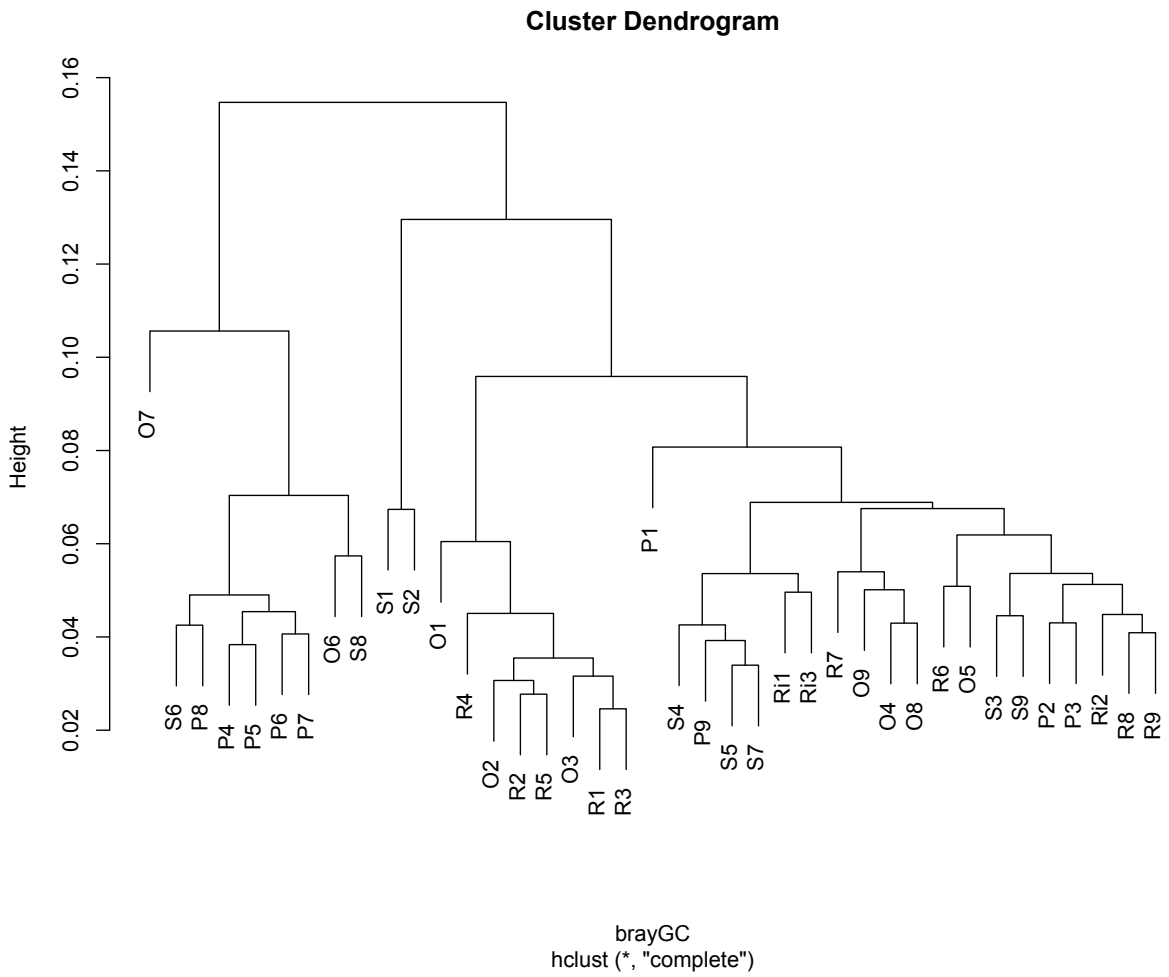


Figure 6. Cluster dendrogram of Bray-Curtis dissimilarity of the full GeoChip dataset.

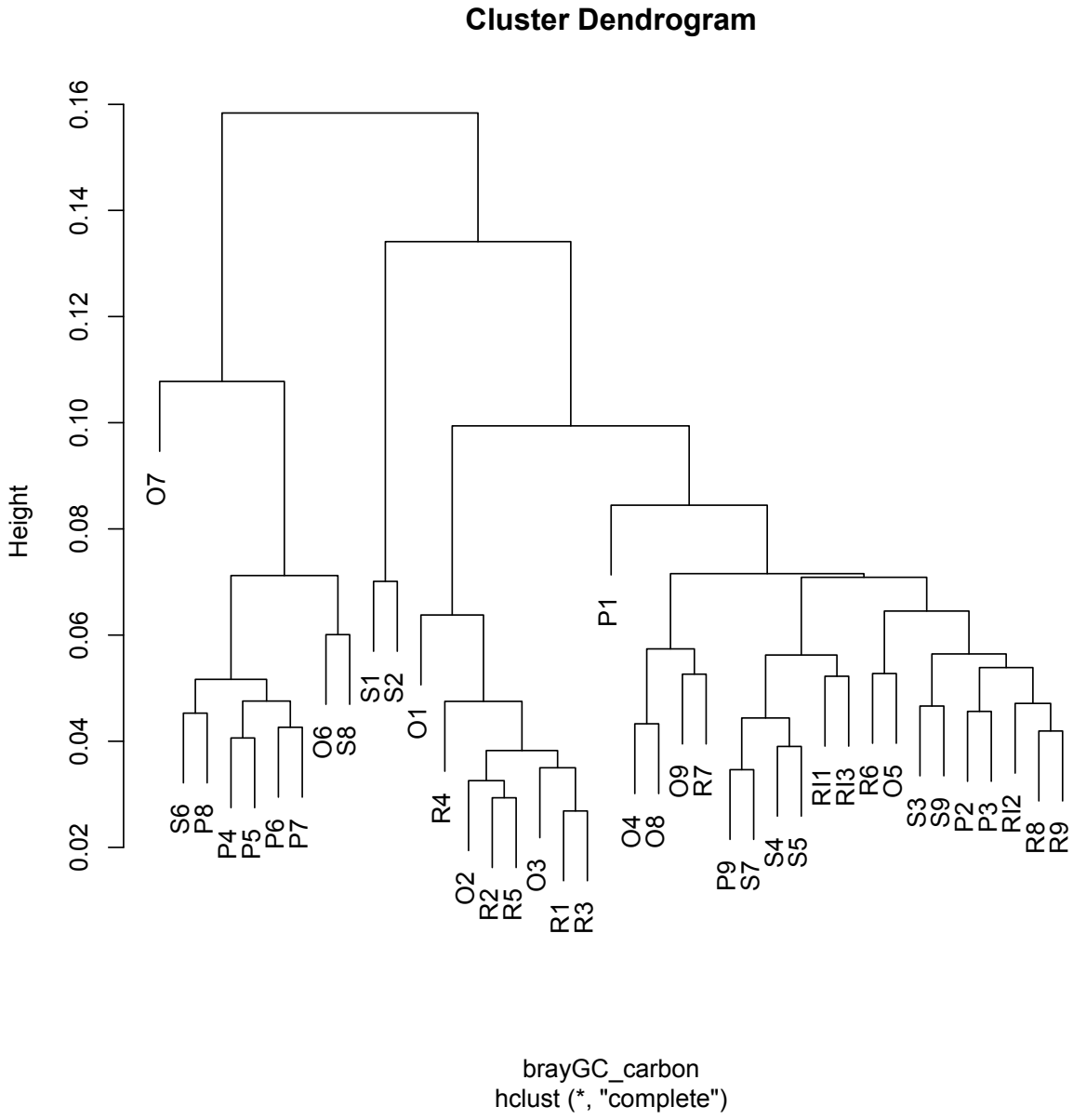


Figure 7. Cluster dendrogram of Bray-Curtis dissimilarity of the GeoChip carbon cycling subset dataset.

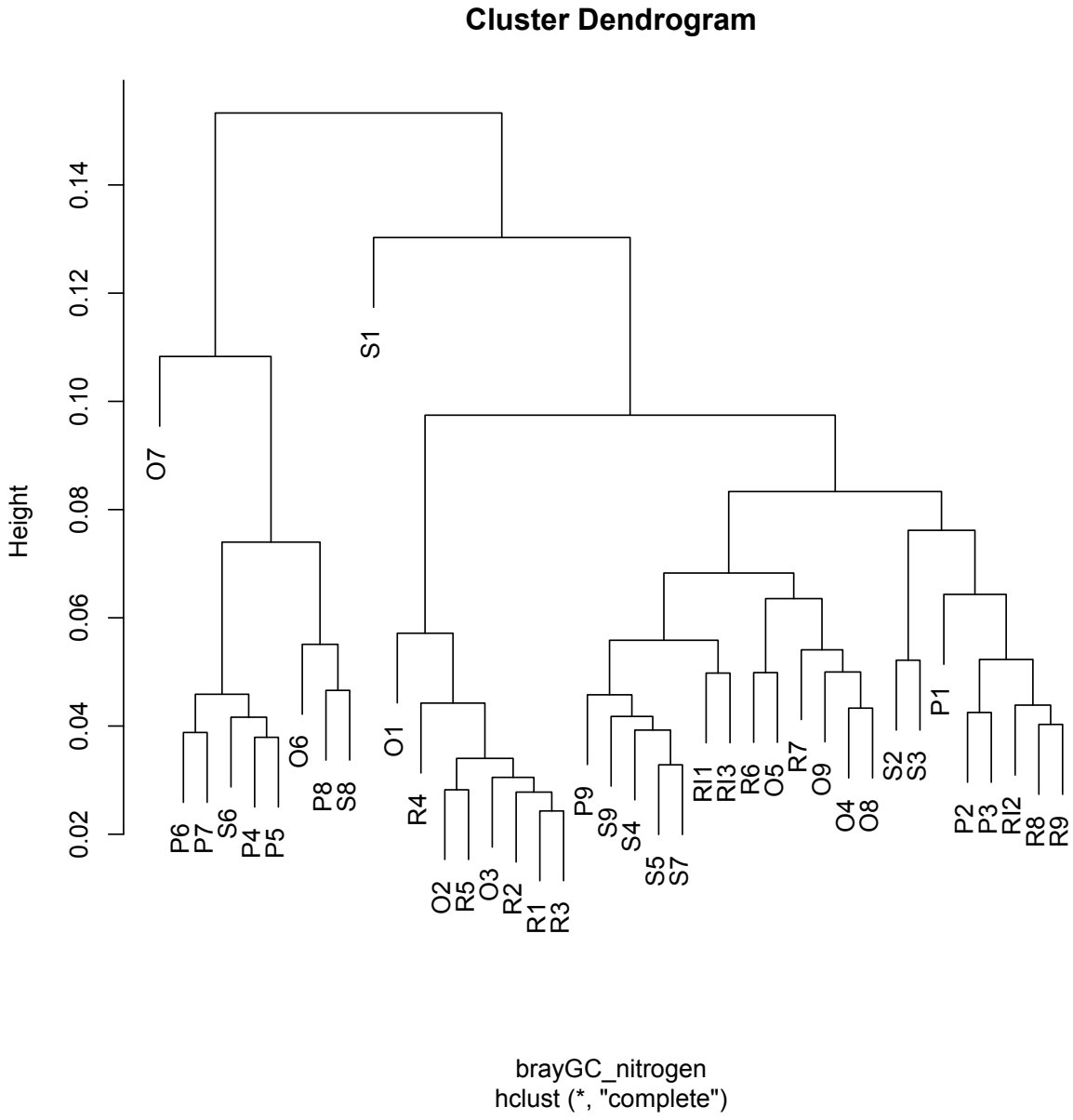


Figure 8. Cluster dendrogram of Bray-Curtis dissimilarity of the GeoChip nitrogen subset dataset.

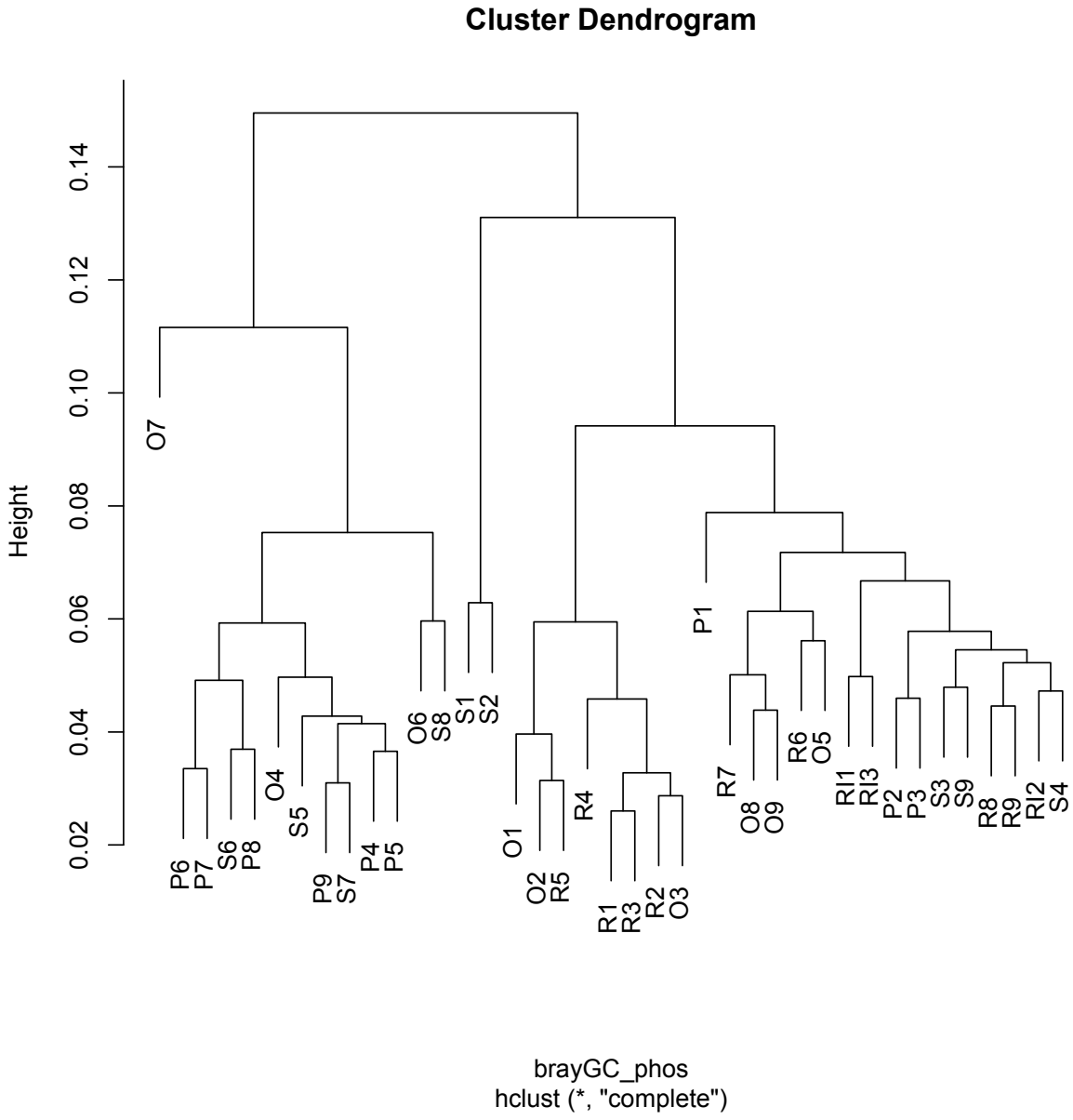


Figure 9. Cluster dendrogram of Bray-Curtis dissimilarity of the GeoChip phosphorus subset dataset.

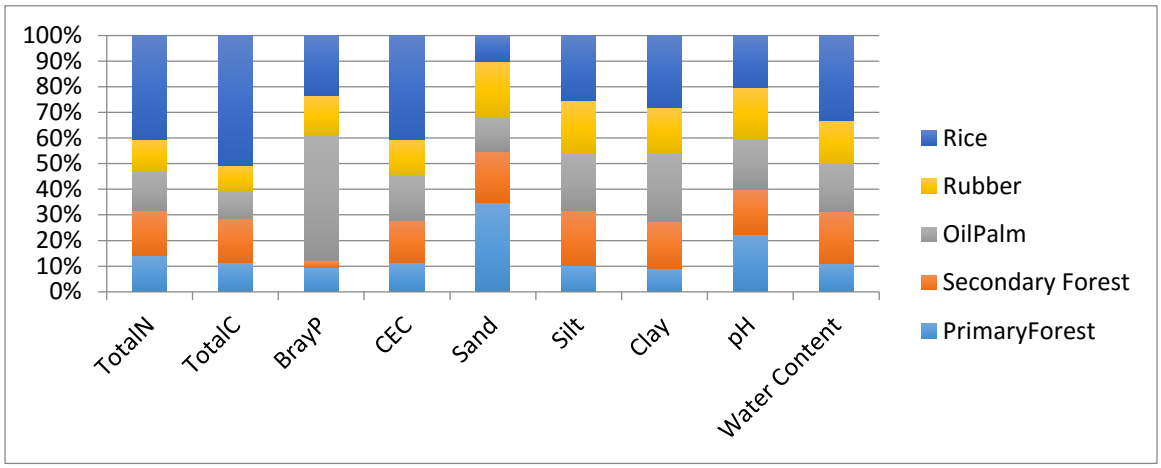


Figure 10. Soil environment variables by land-use type.

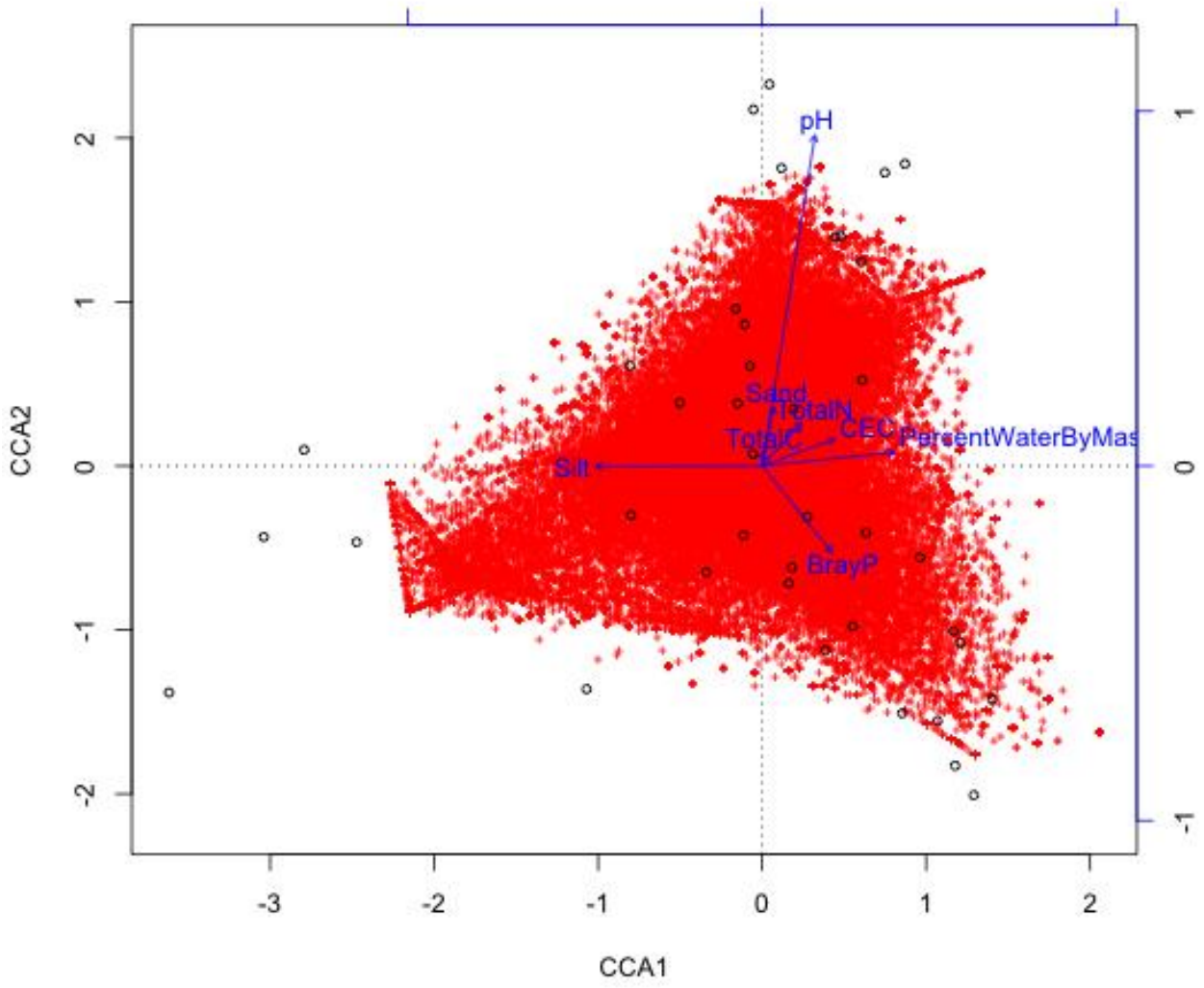


Figure 11. CCA ordination with environmental variables of the 16S dataset.

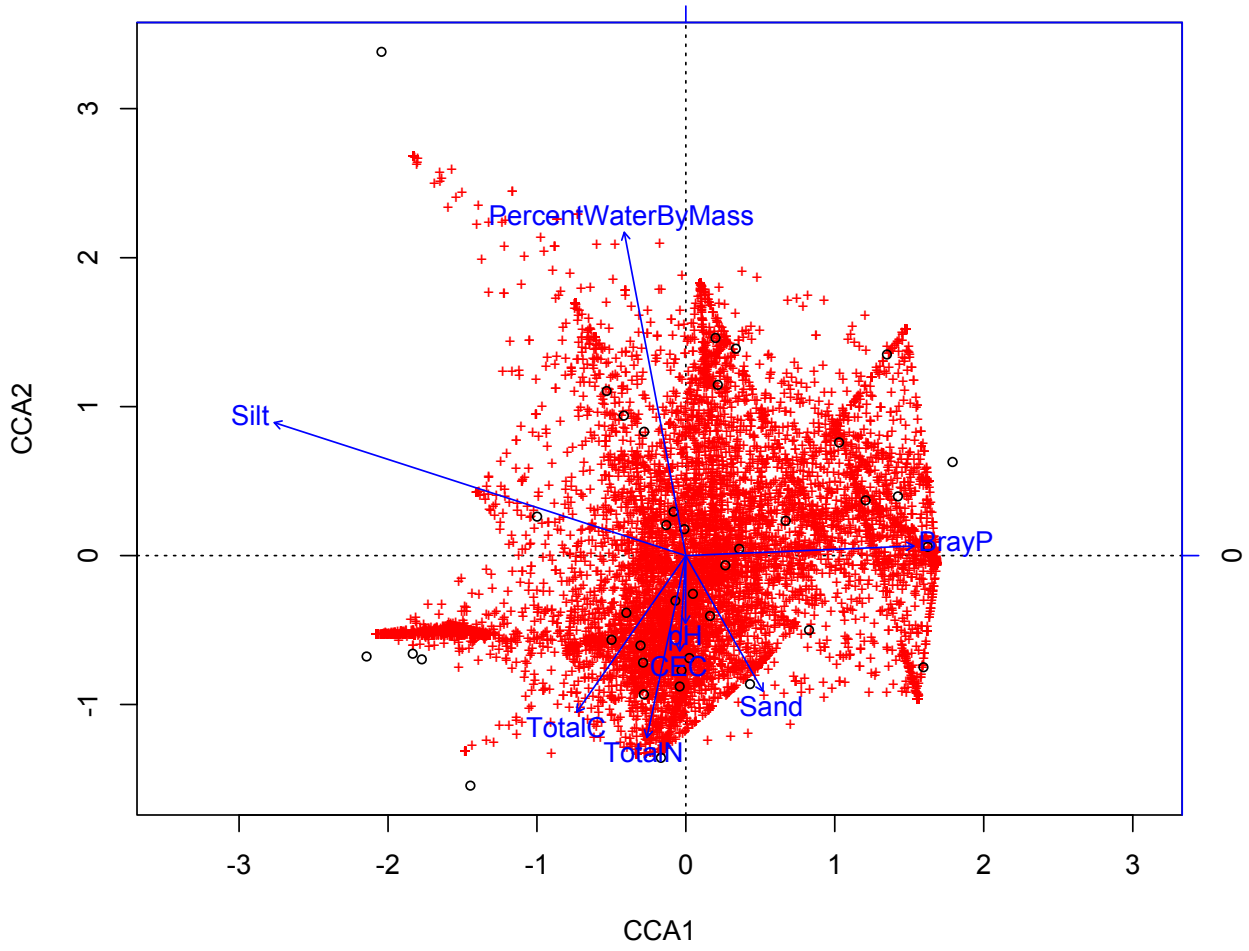


Figure 12. CCA ordination with environmental variables of the ITS dataset.

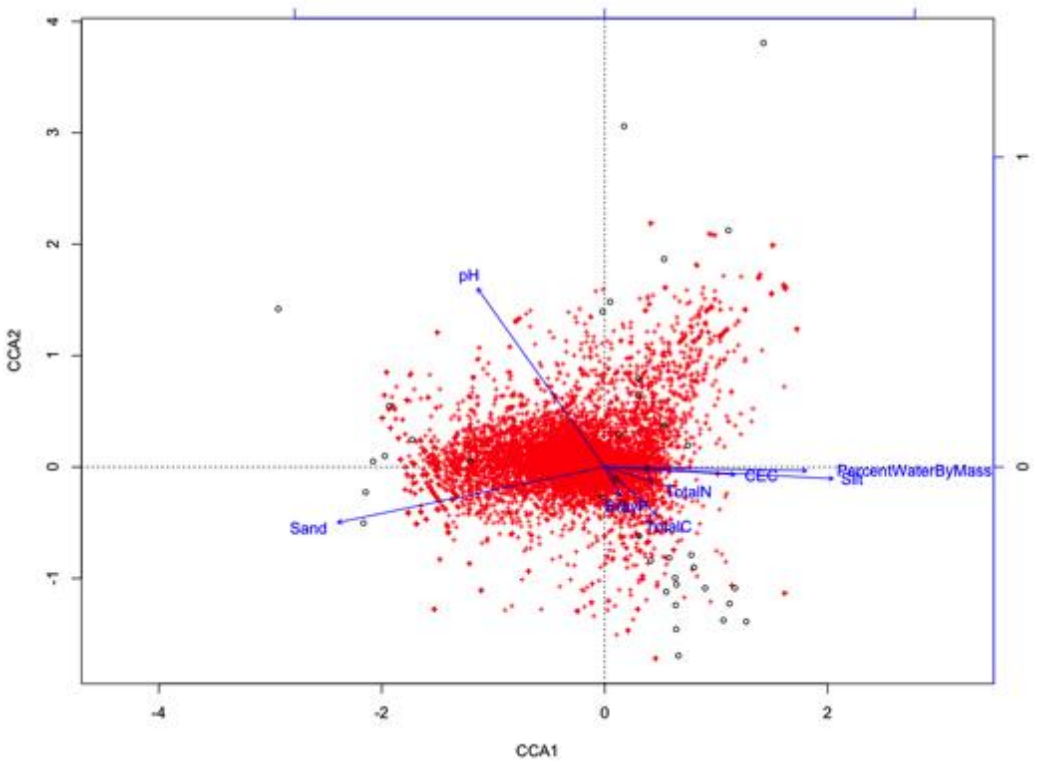


Figure 13. CCA ordination with environmental variables of the full GeoChip dataset.

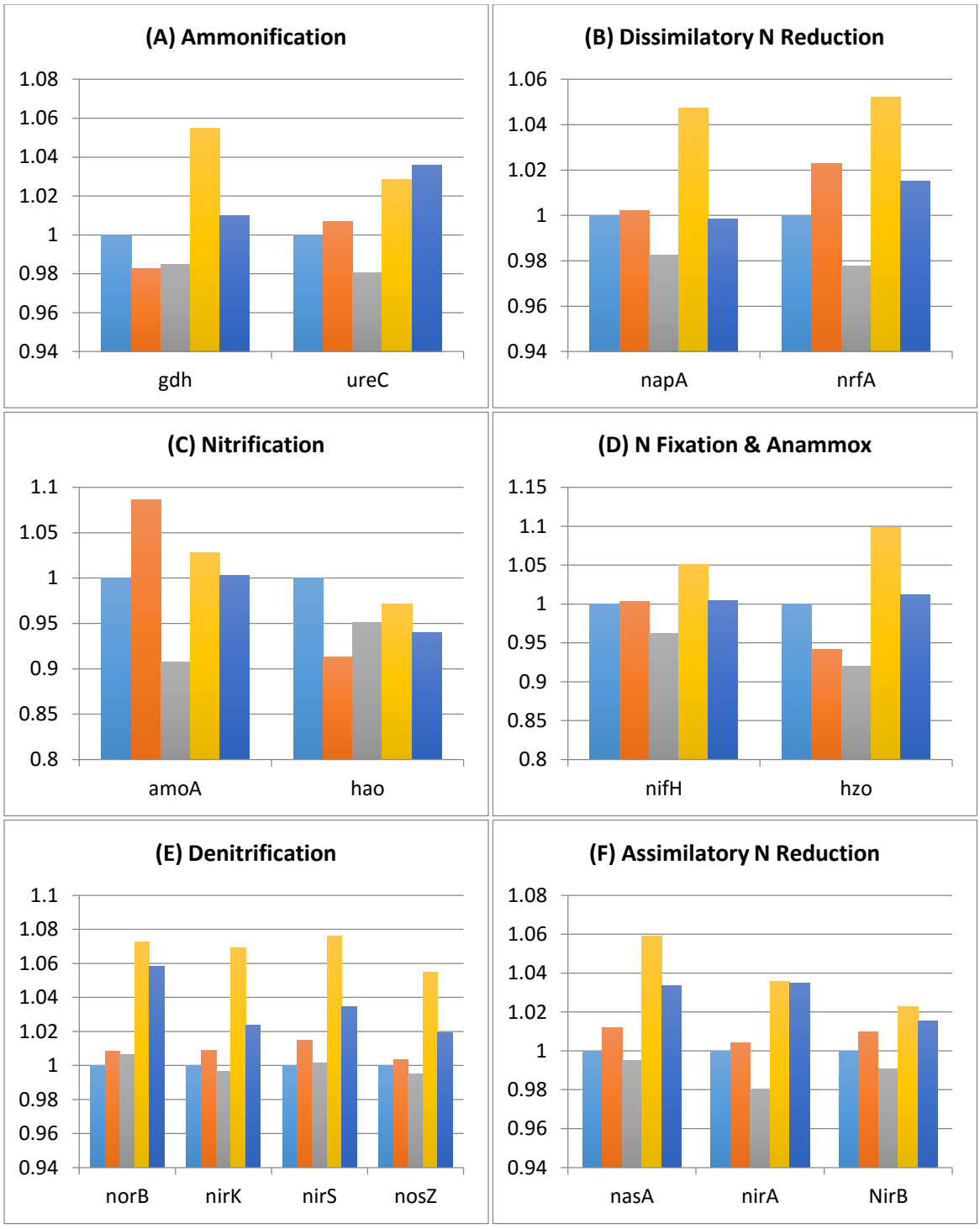


Figure 14. Total gene abundances, normalized by Primary Forest abundances, of Nitrogen cycling genes in each land-use type. The plotted data are means of the nine sampling plots per land-use type (or three plots in the case of Rice). The y-axes are normalized total gene abundances. Primary Forest, Secondary Forest, Oil Palm, Rubber, Rice

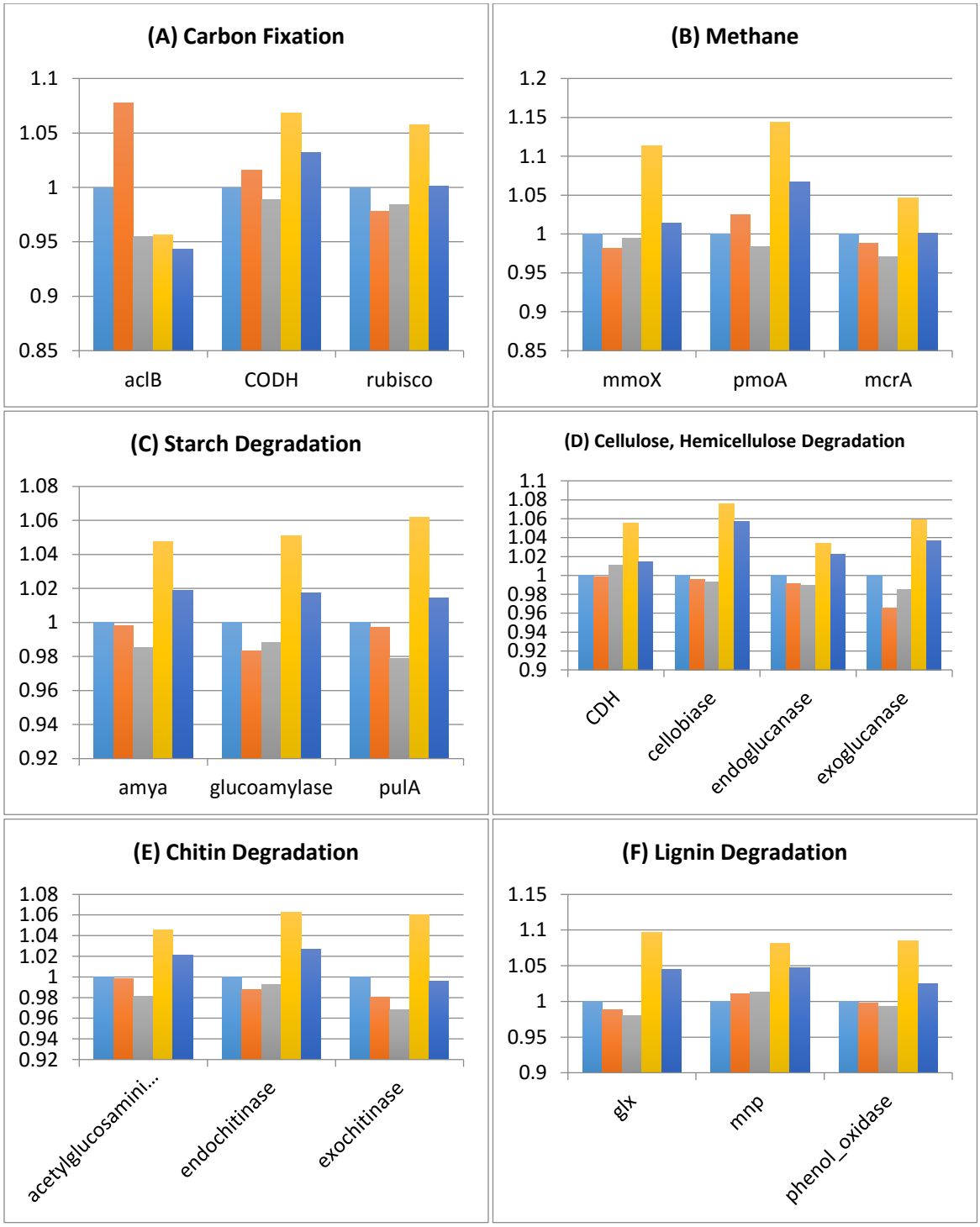


Figure 15. Total gene abundances, normalized by Primary Forest abundances, of Carbon cycling genes in each land-use type. The plotted data are means of the nine sampling plots per land-use type (or three plots in the case of Rice). The y-axes are normalized total gene abundances. Primary Forest, Secondary Forest, Oil Palm, Rubber, Rice

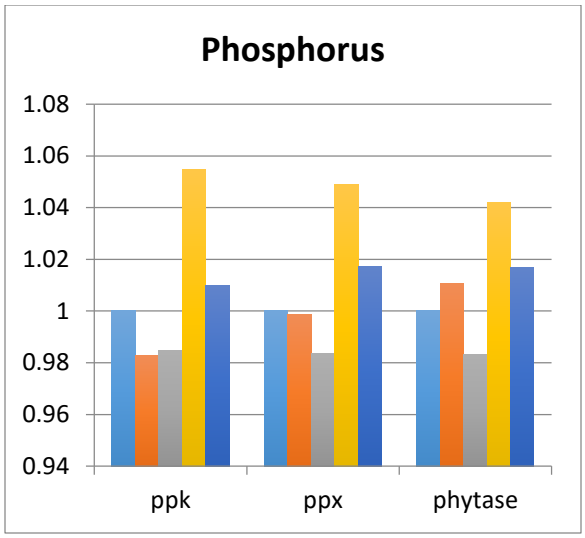
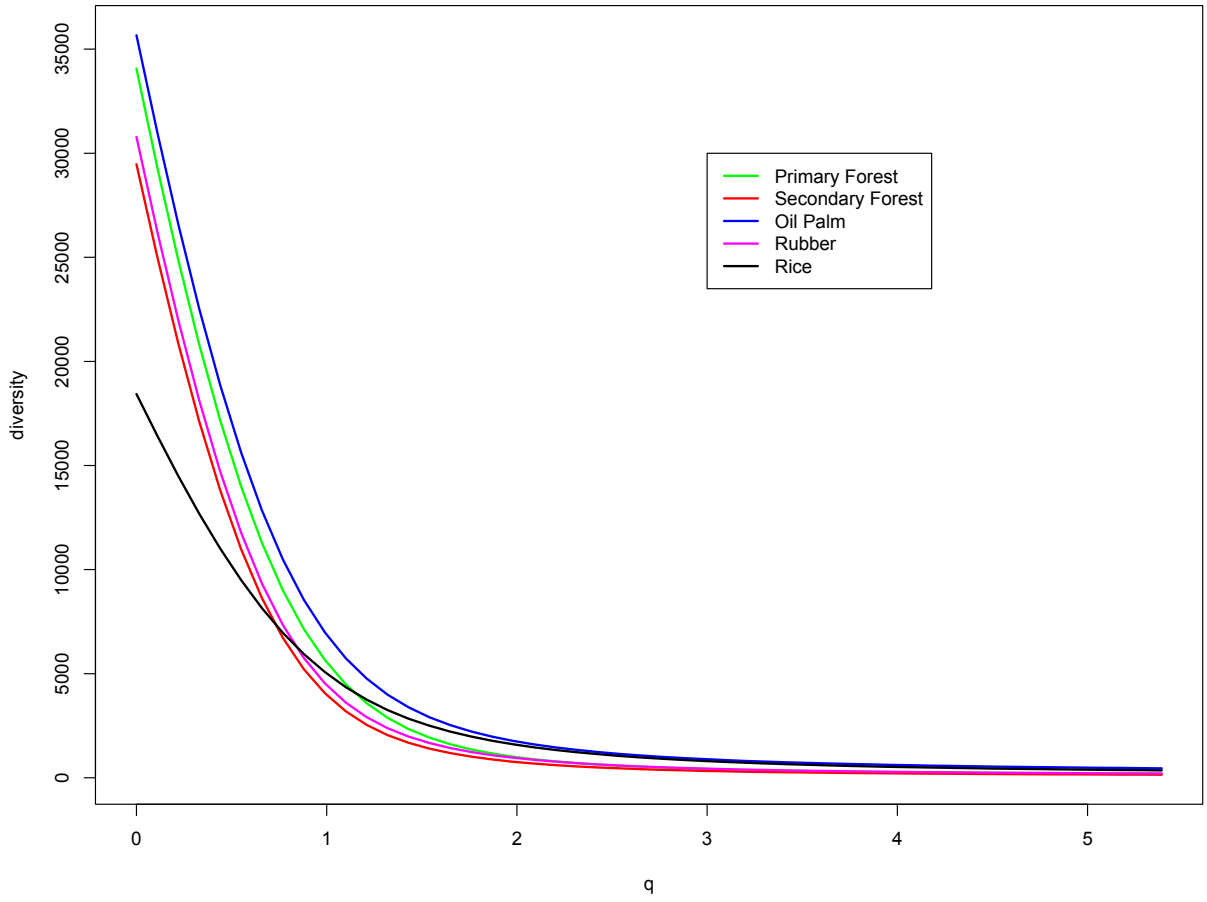


Figure 16. Total gene abundances, normalized by Primary Forest abundances, of Phosphorus cycling genes in each land-use type. The plotted data are means of the nine sampling plots per land-use type (or three plots in the case of Rice). The y-axis is normalized total gene abundances. Primary Forest, Secondary Forest, Oil Palm, Rubber, Rice

Supplementary Material

(A)



(B)

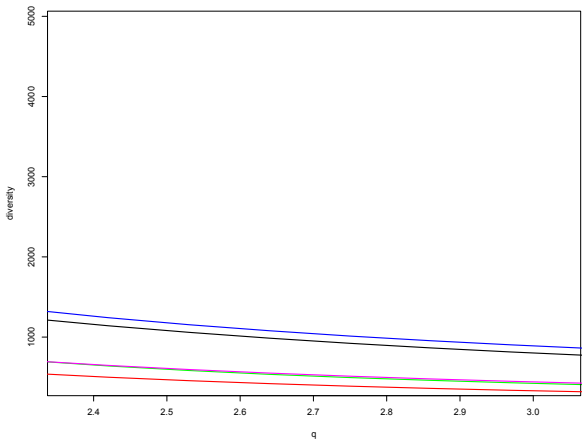


Figure S1. (A) Diversity profile calculated for the full 16S dataset, including the outlier sample, O1. (B) A closer view of the portion of the diversity profile where $2.4 \leq q \leq 3.8$.

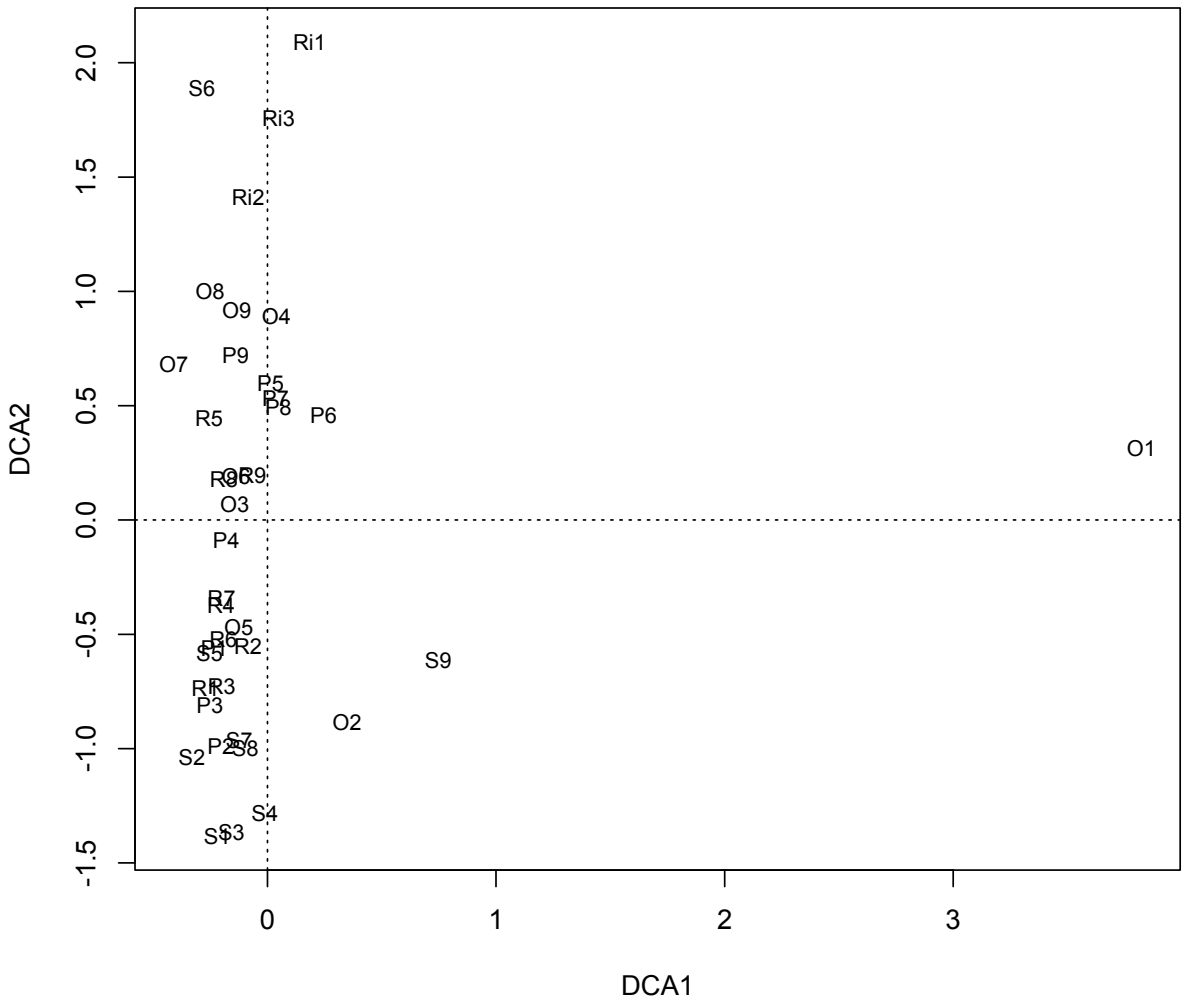


Figure S2. DCA ordination of the full 16S dataset, including the outlier sample, O1.

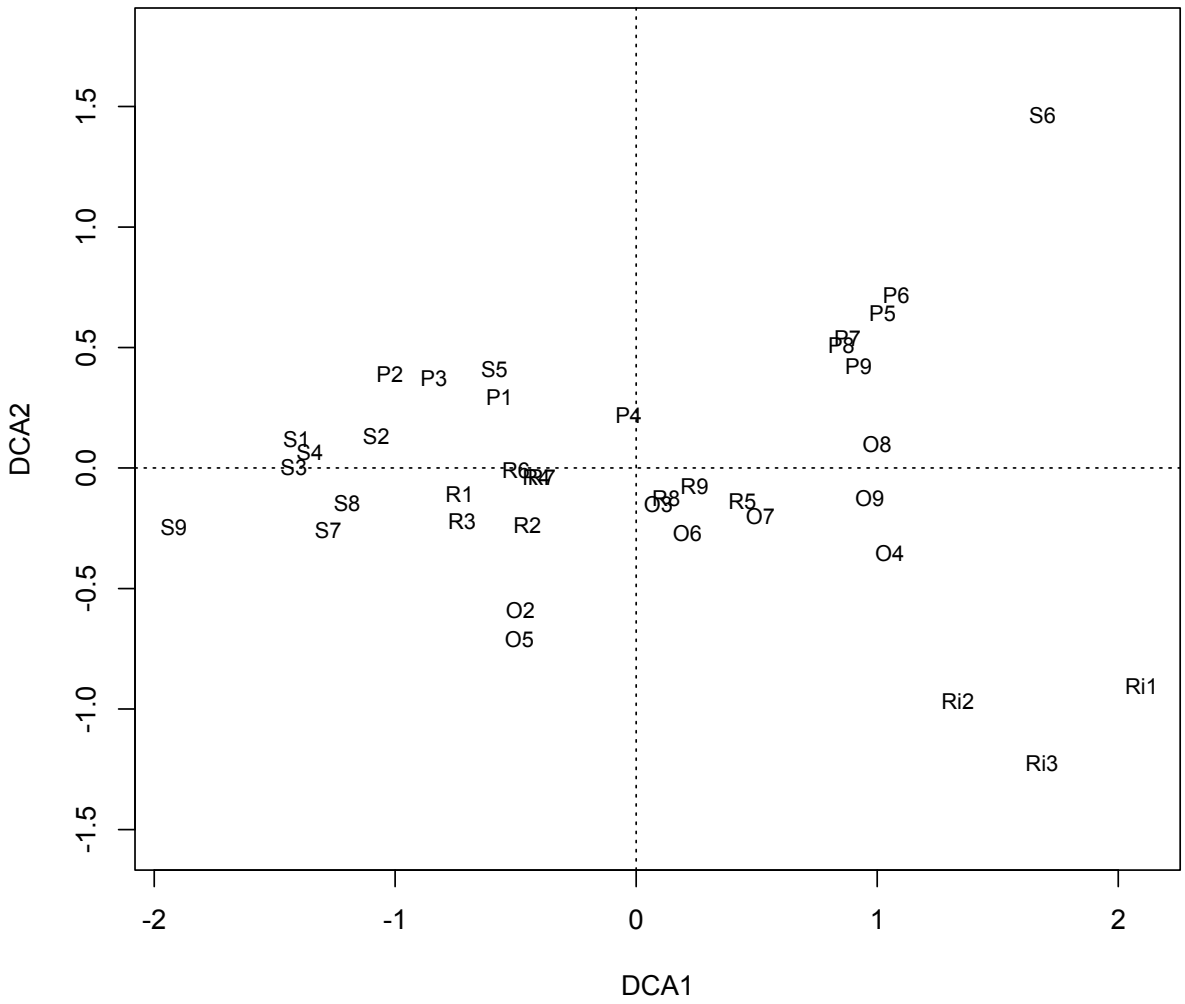


Figure S3. DCA ordination of the 16S dataset, without the outlier sample, O1.

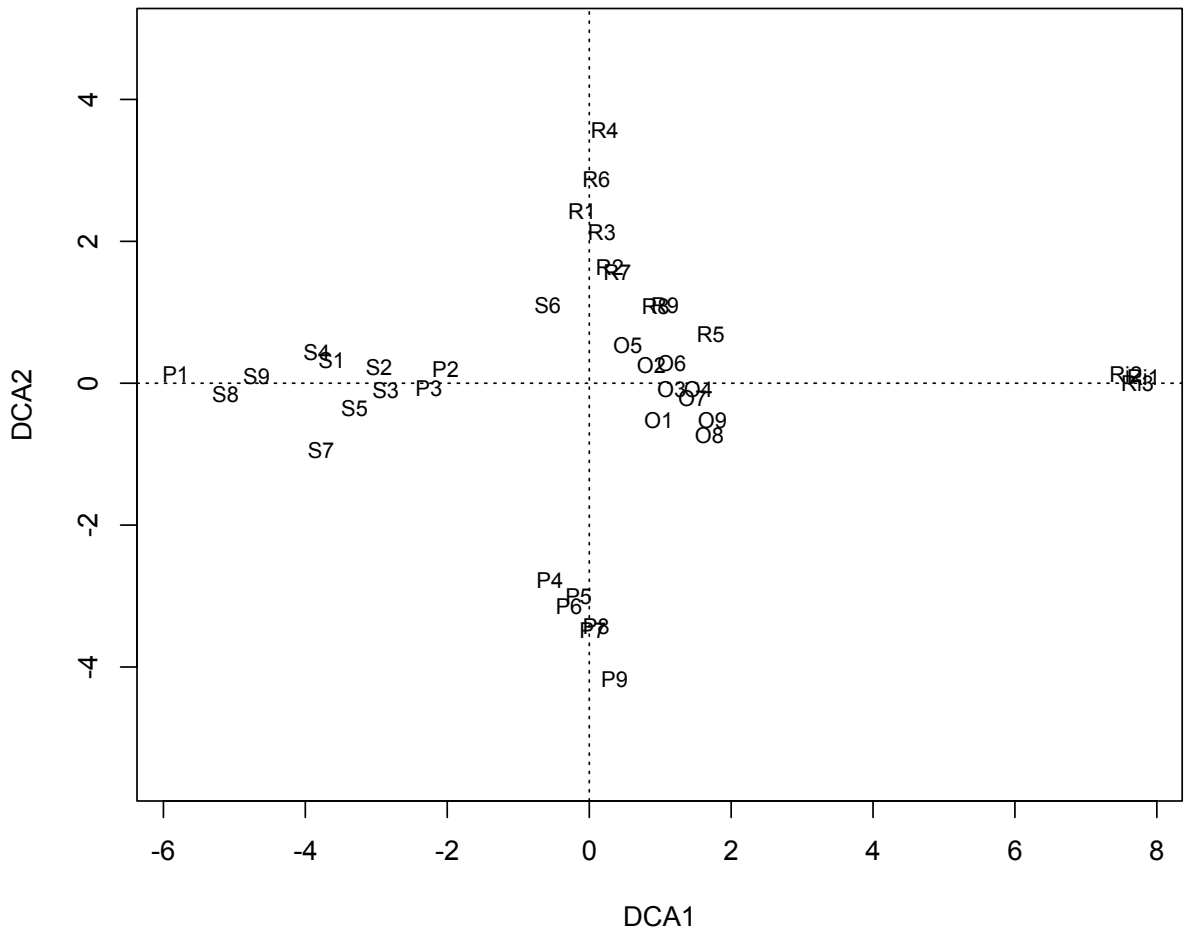


Figure S4. DCA ordination of the full ITS dataset.

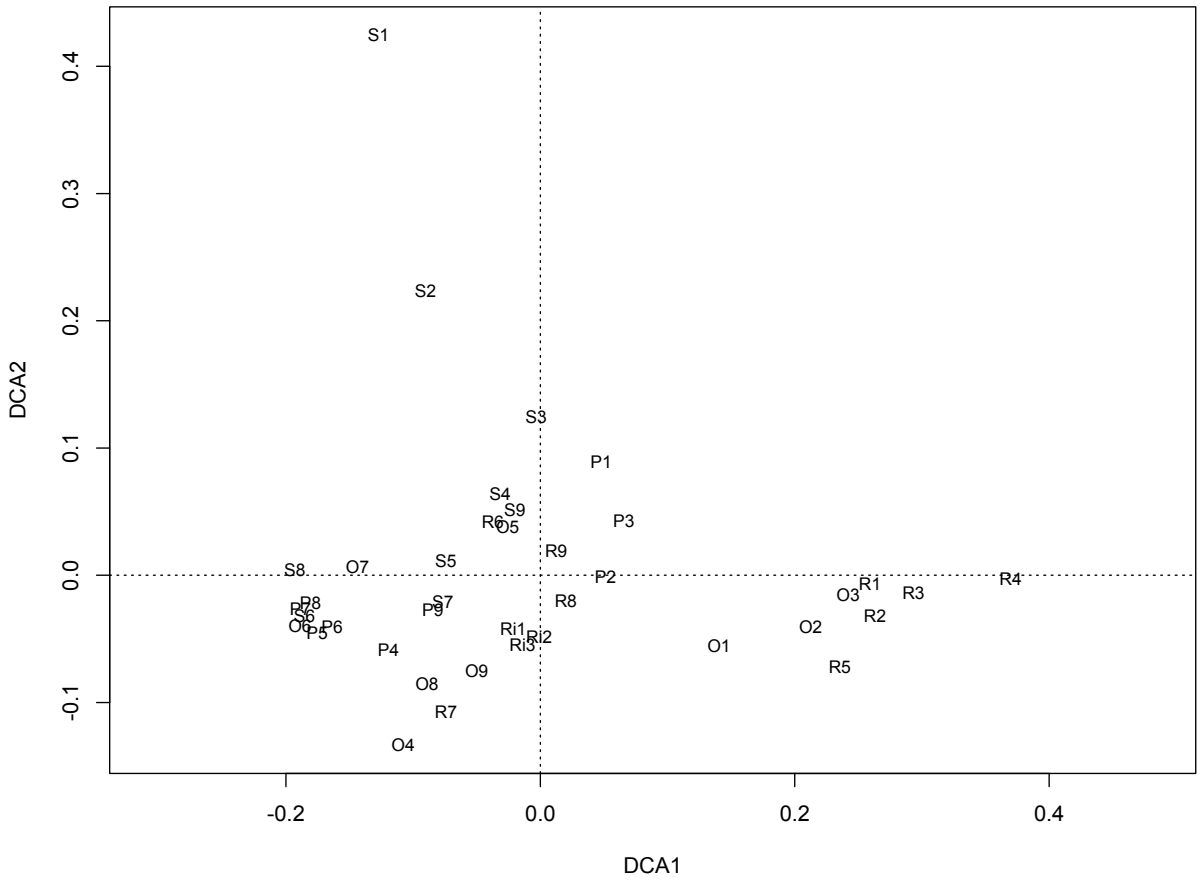


Figure S5. DCA ordination of the full GeoChip dataset.

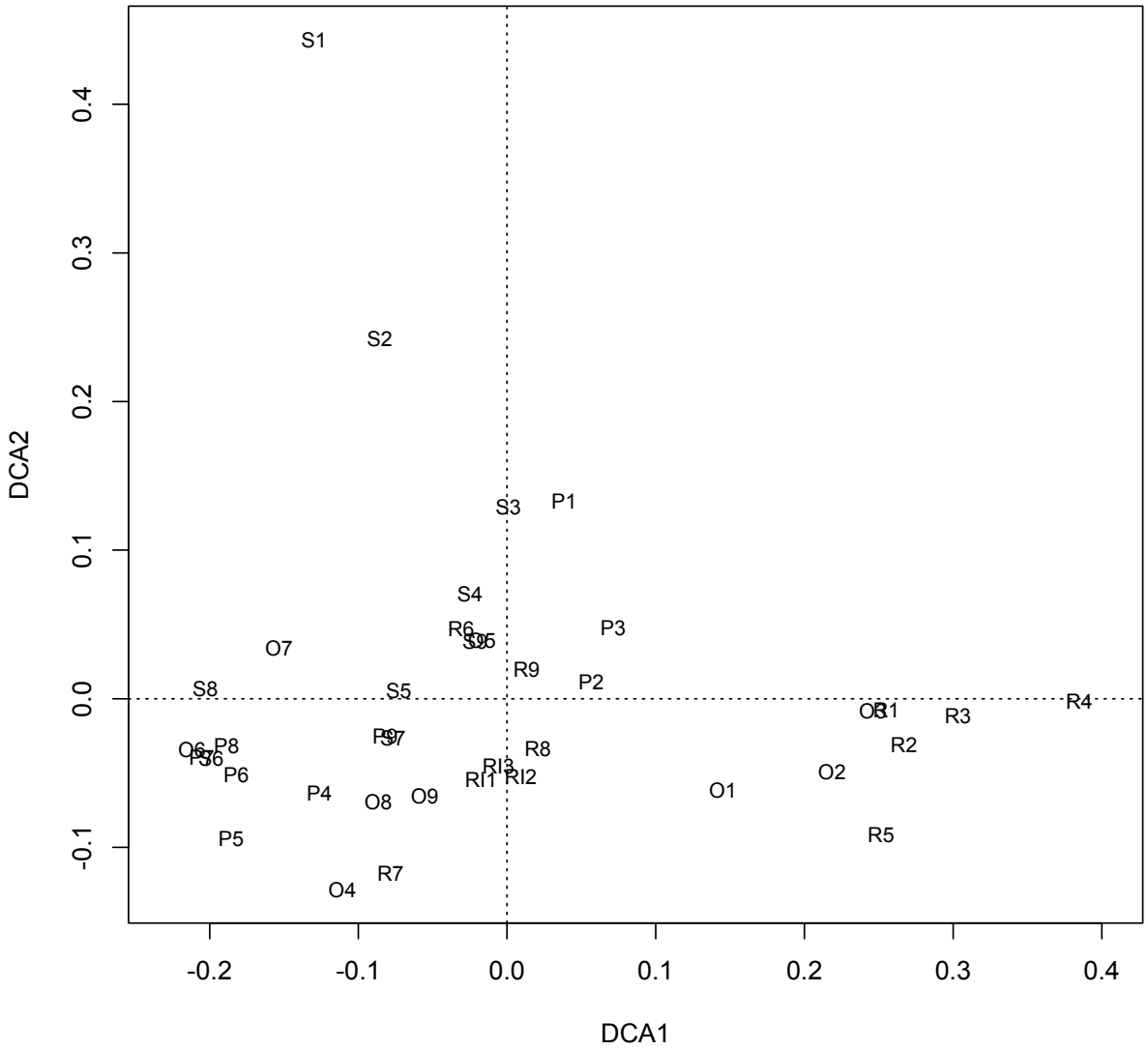


Figure S6. DCA ordination of the GeoChip carbon cycling subset dataset.

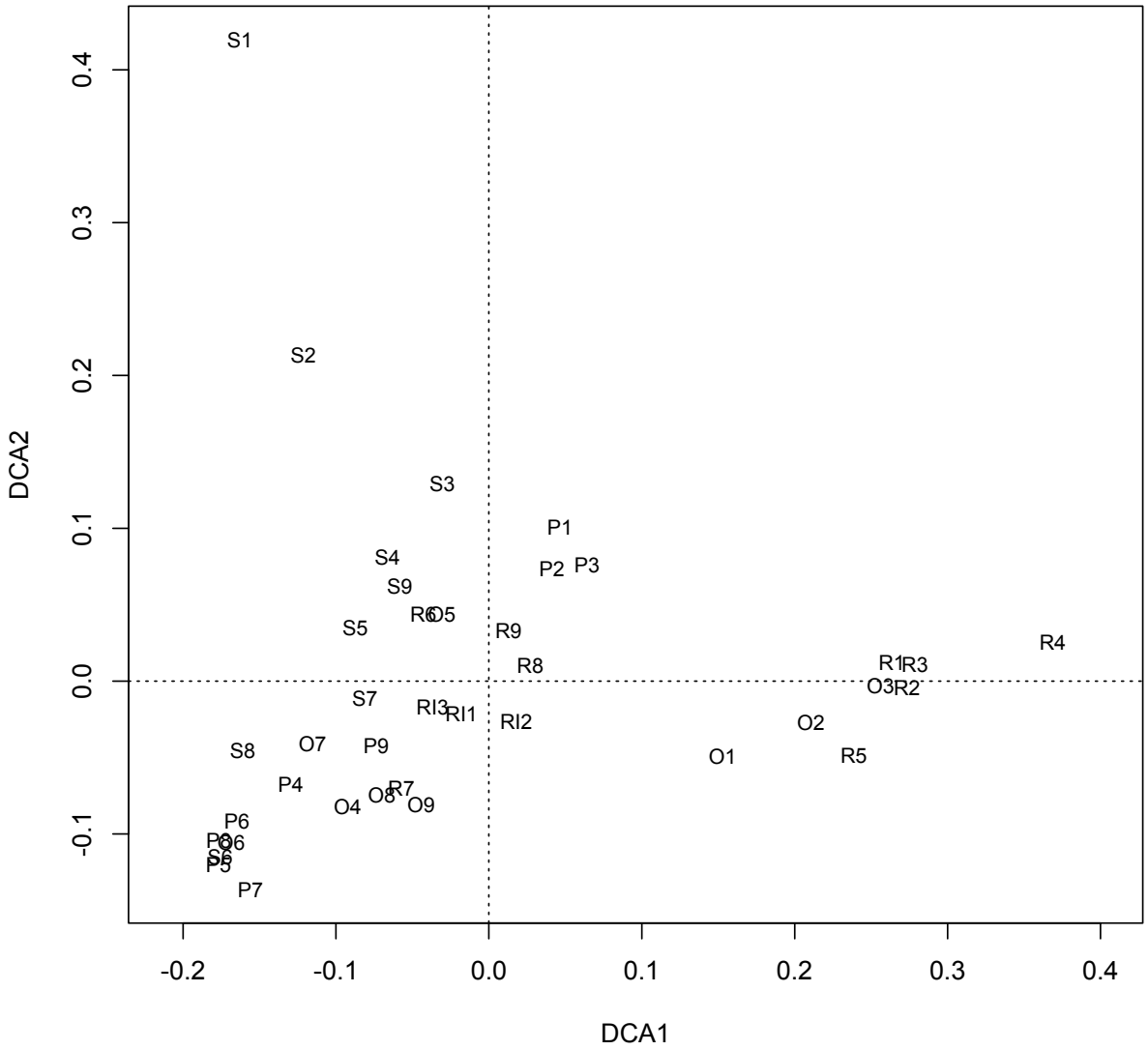


Figure S7. DCA ordination of the GeoChip nitrogen subset dataset.

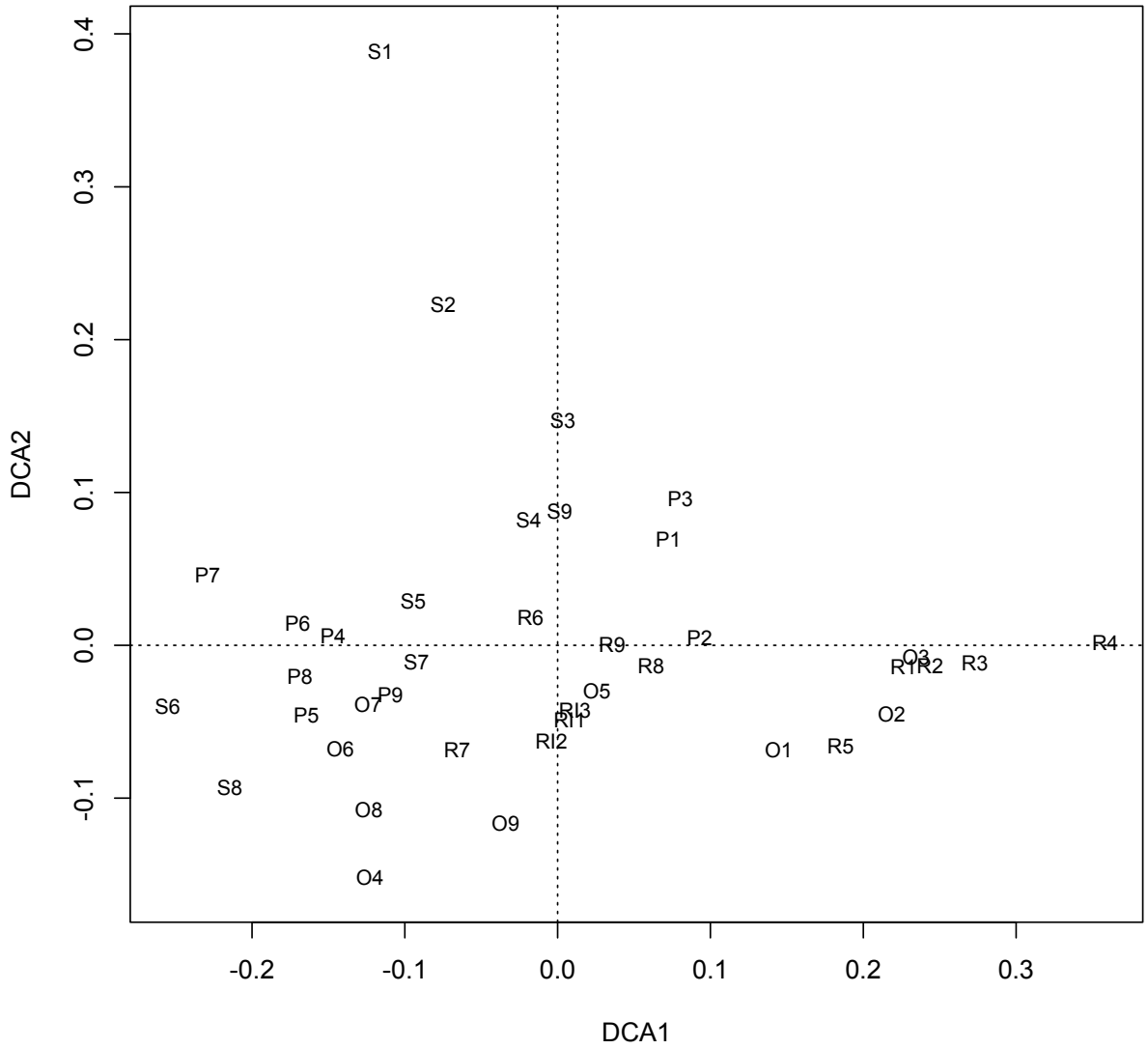


Figure S8. DCA ordination of the GeoChip phosphorus subset dataset.

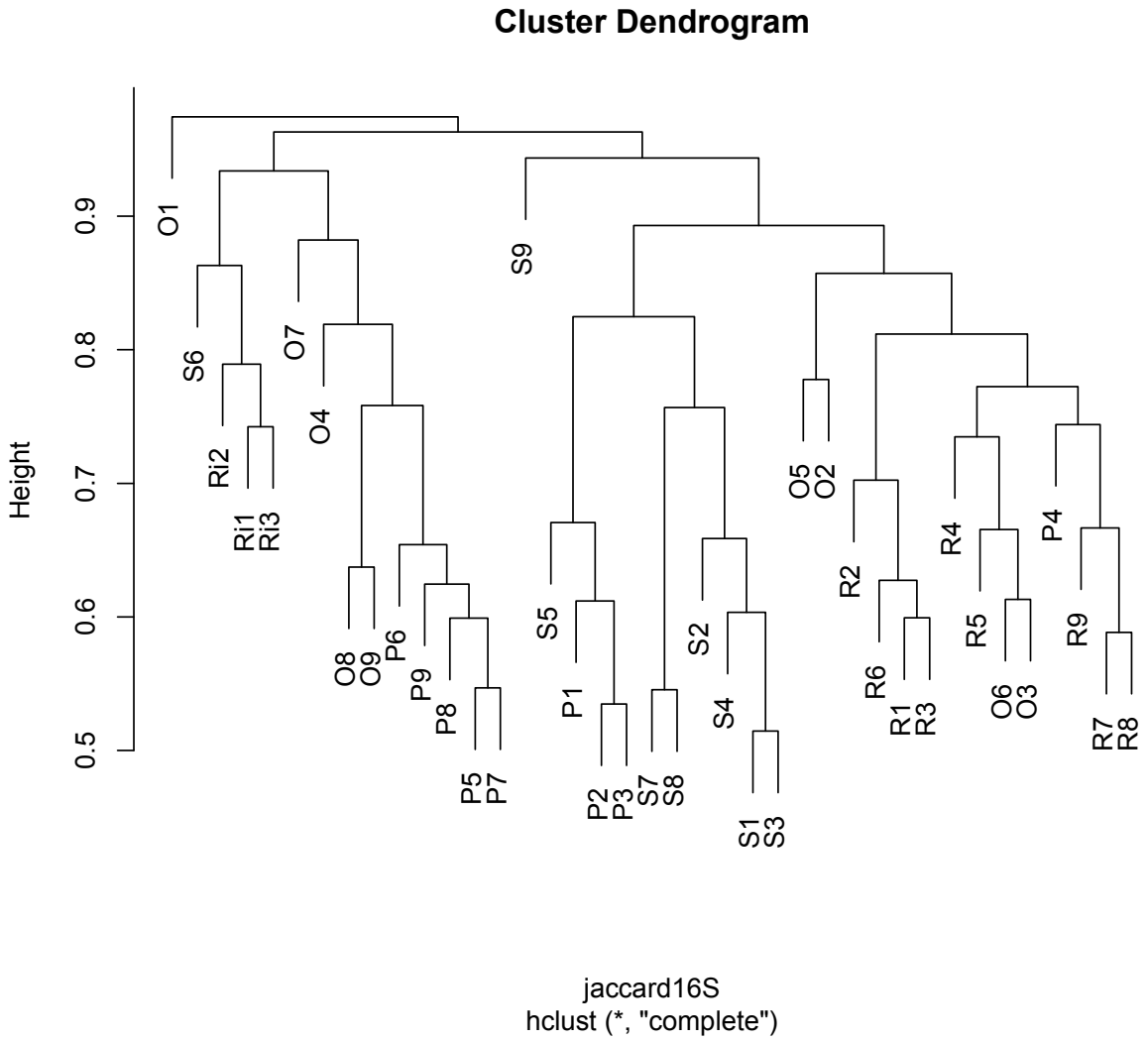


Figure S9. Cluster dendrogram of Jaccard diversity of the 16S dataset.

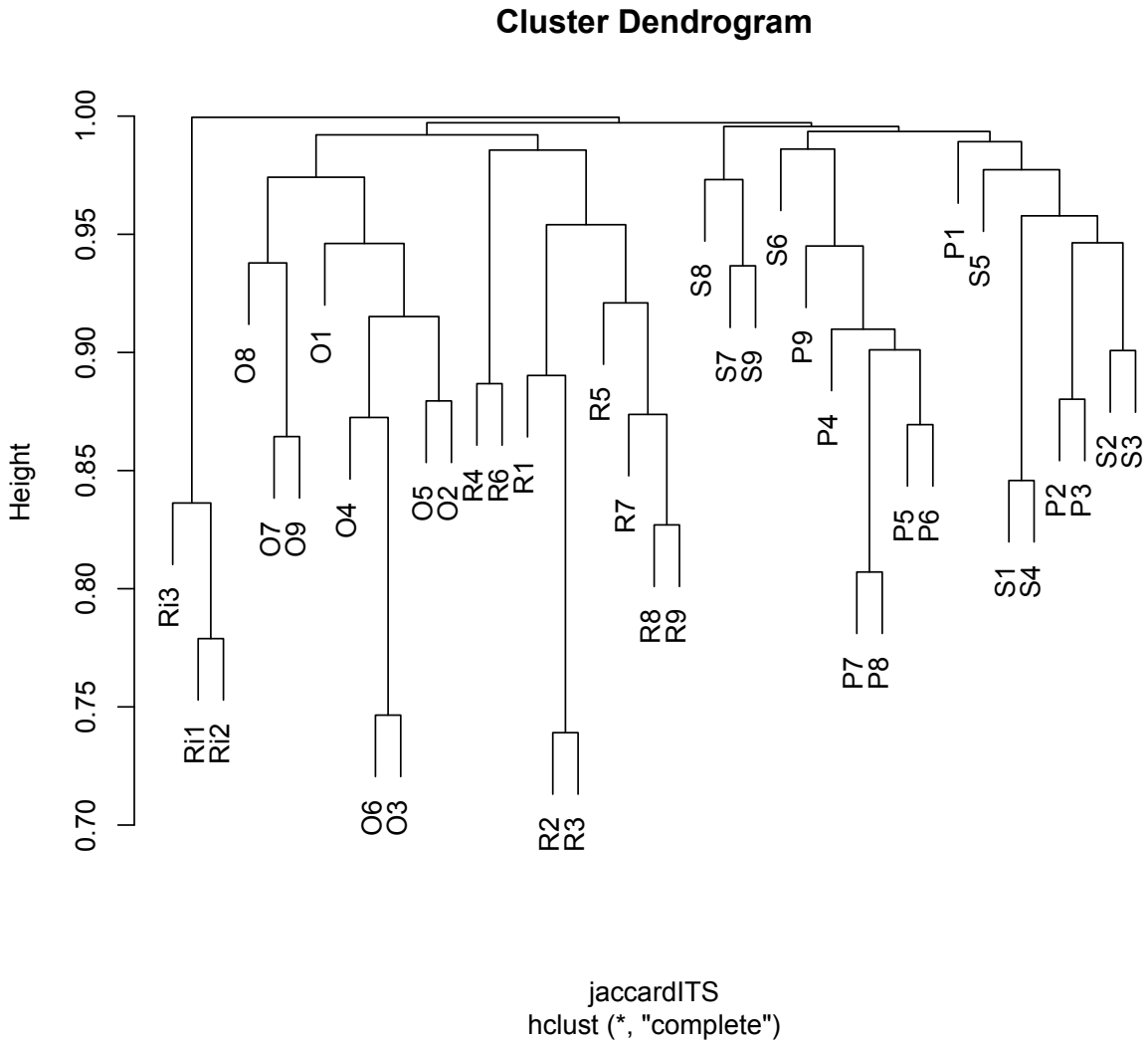


Figure S10. Cluster dendrogram of Jaccard diversity of the ITS dataset.

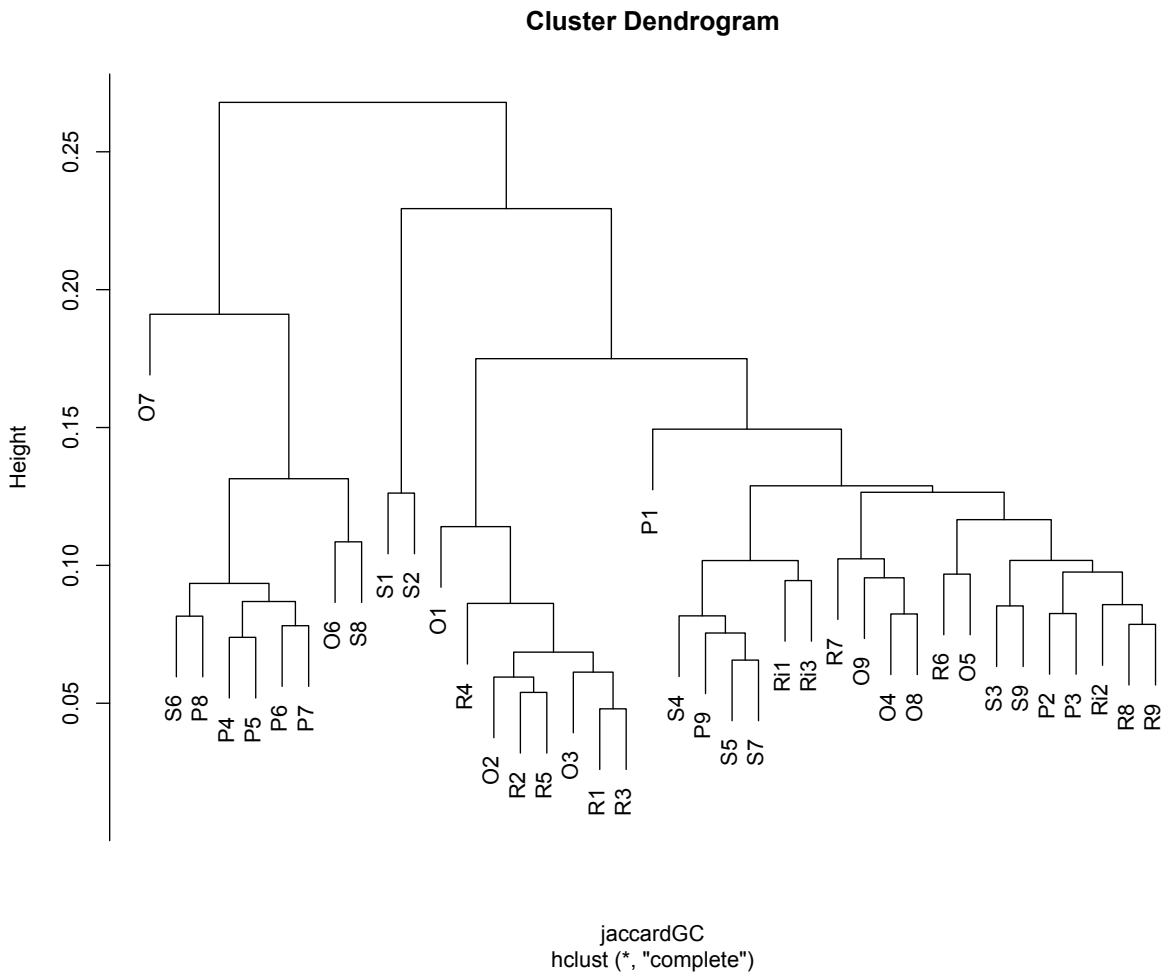


Figure S11. Cluster dendrogram of Jaccard diversity of the full GeoChip dataset.

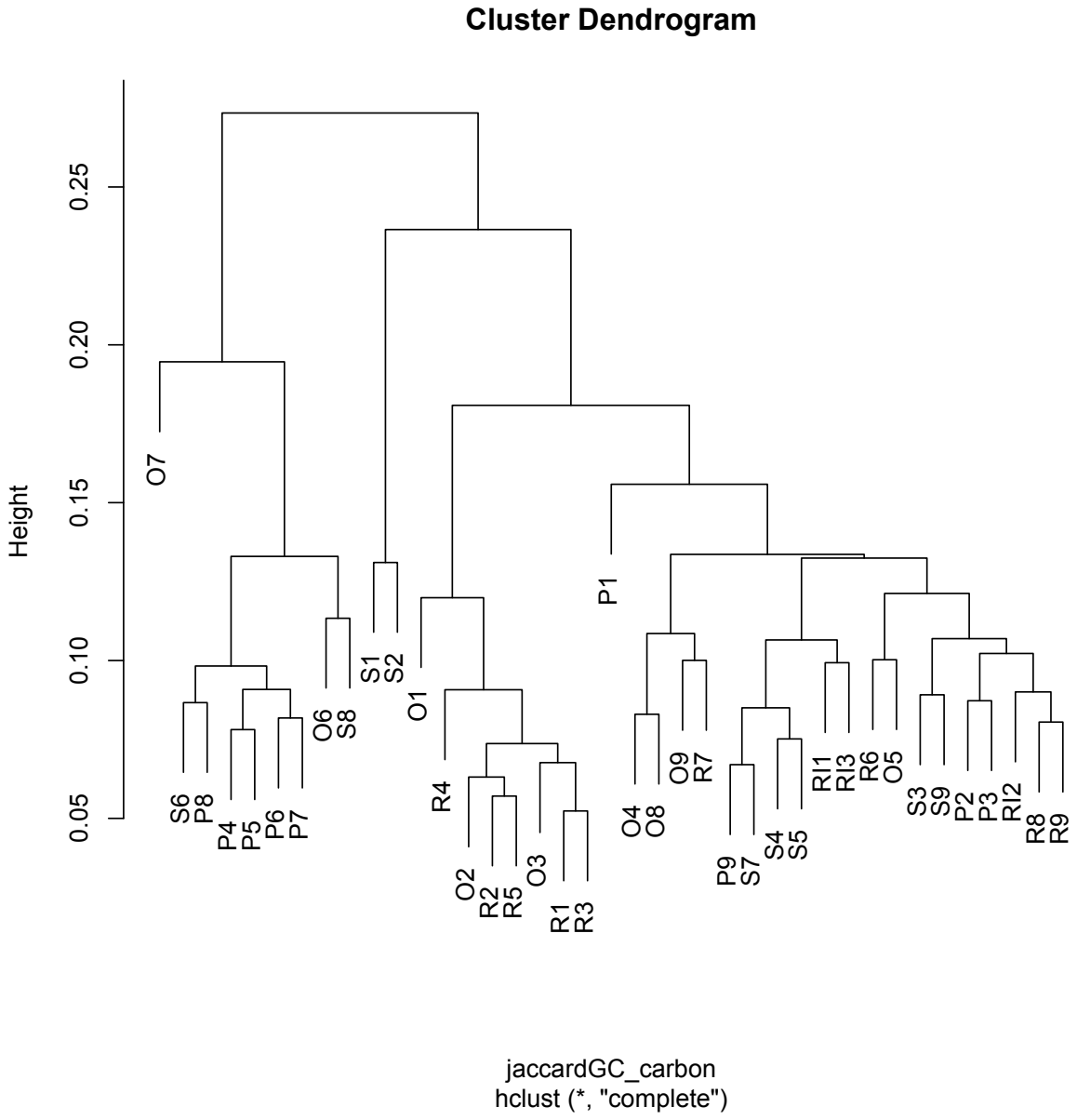


Figure S12. Cluster dendrogram of Jaccard diversity of the GeoChip carbon cycling subset dataset.

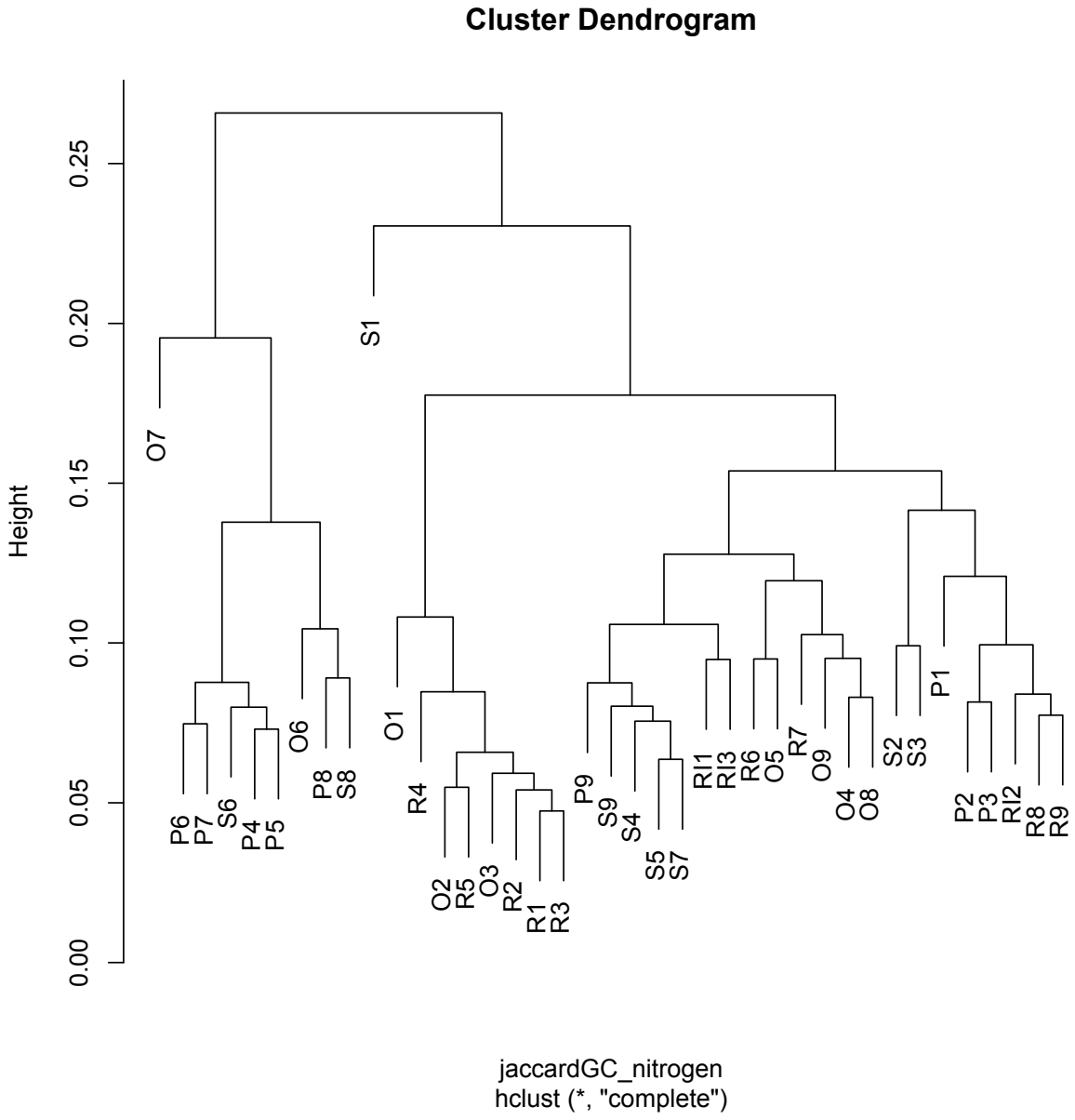


Figure S13. Cluster dendrogram of Jaccard diversity of the GeoChip nitrogen subset dataset.

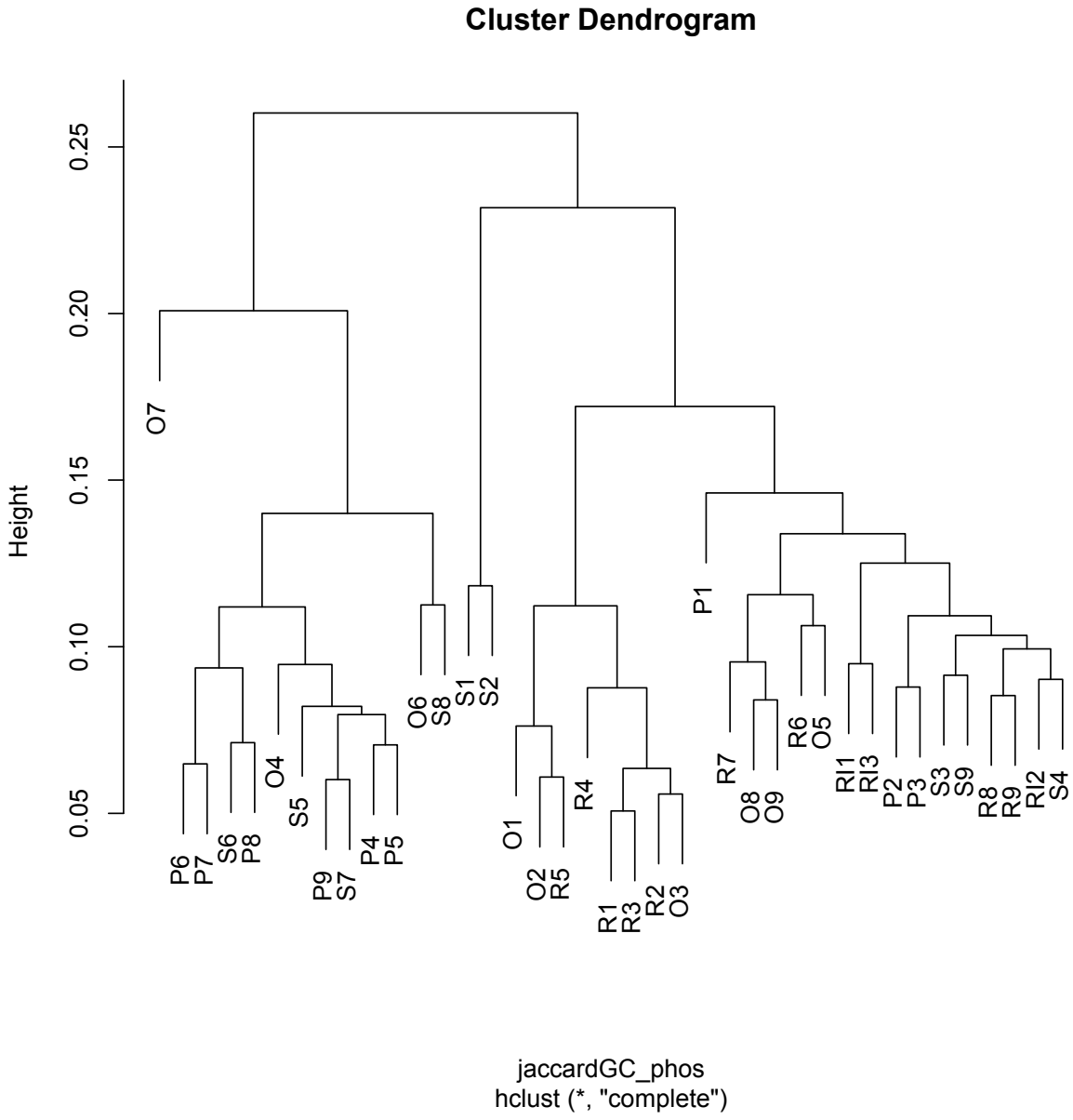


Figure S14. Cluster dendrogram of Jaccard diversity of the GeoChip phosphorus subset dataset.

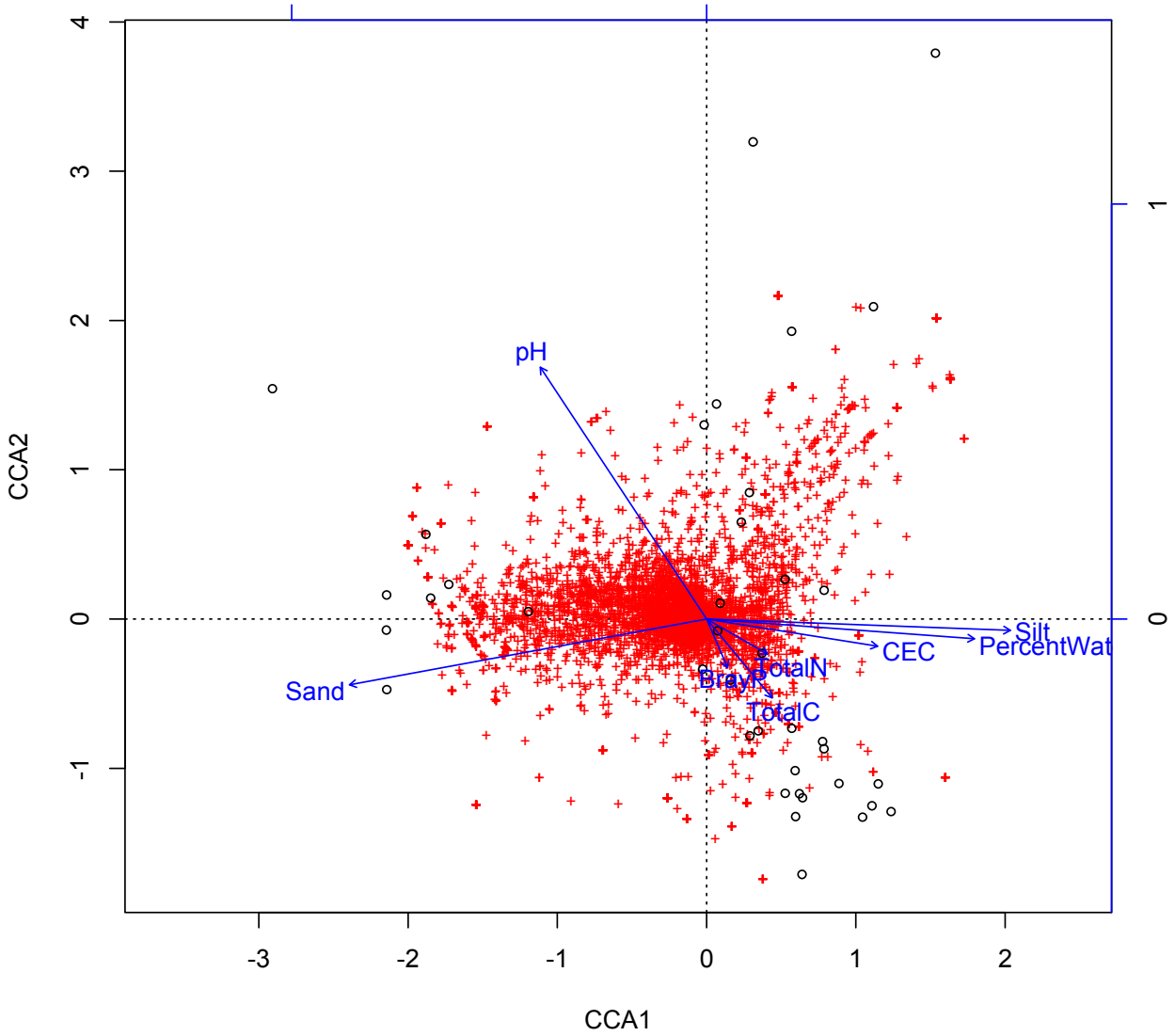


Figure S15. CCA ordination with environmental variables of the GeoChip carbon cycling subset dataset.

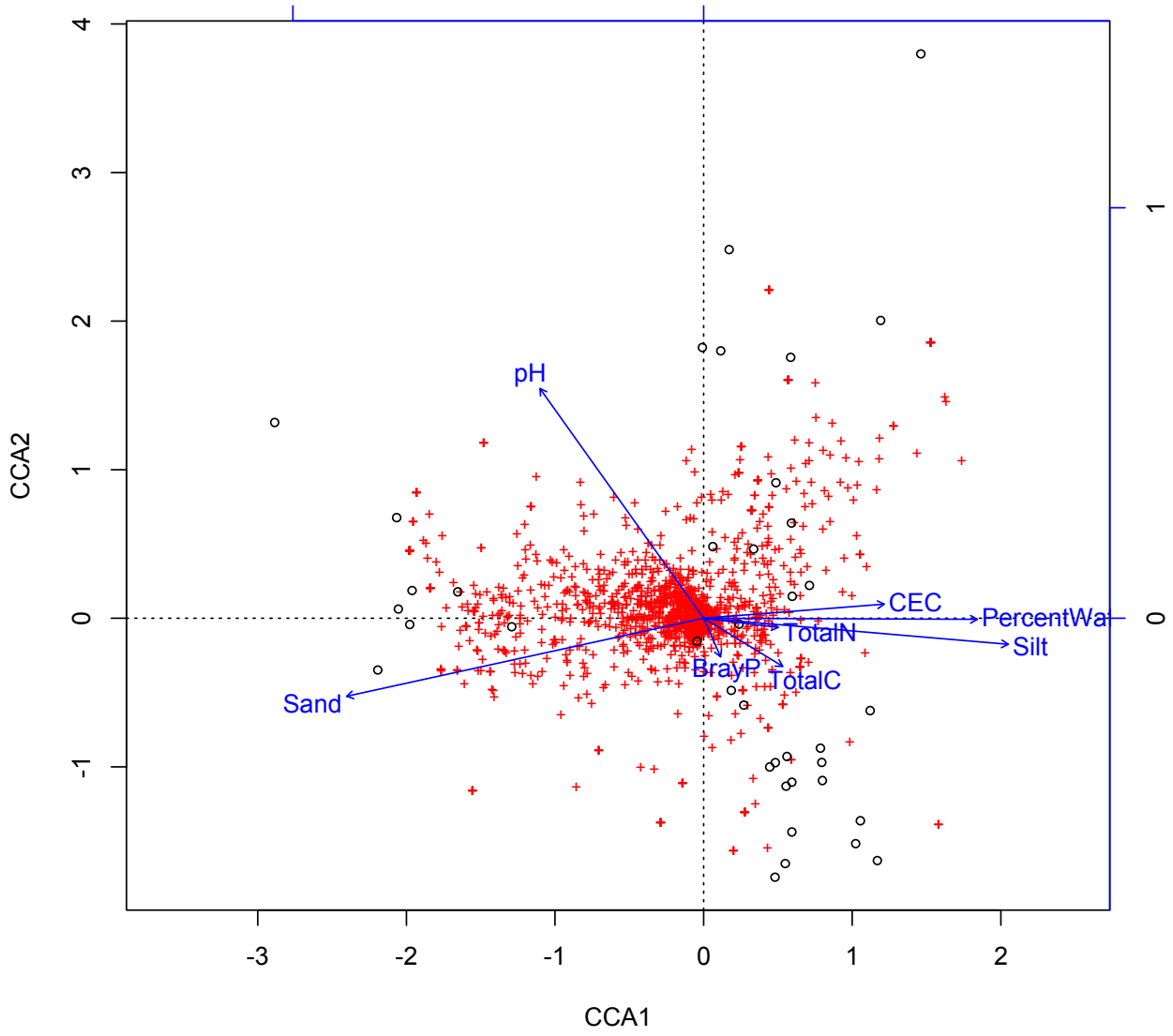


Figure S16. CCA ordination with environmental variables of the GeoChip nitrogen subset dataset.

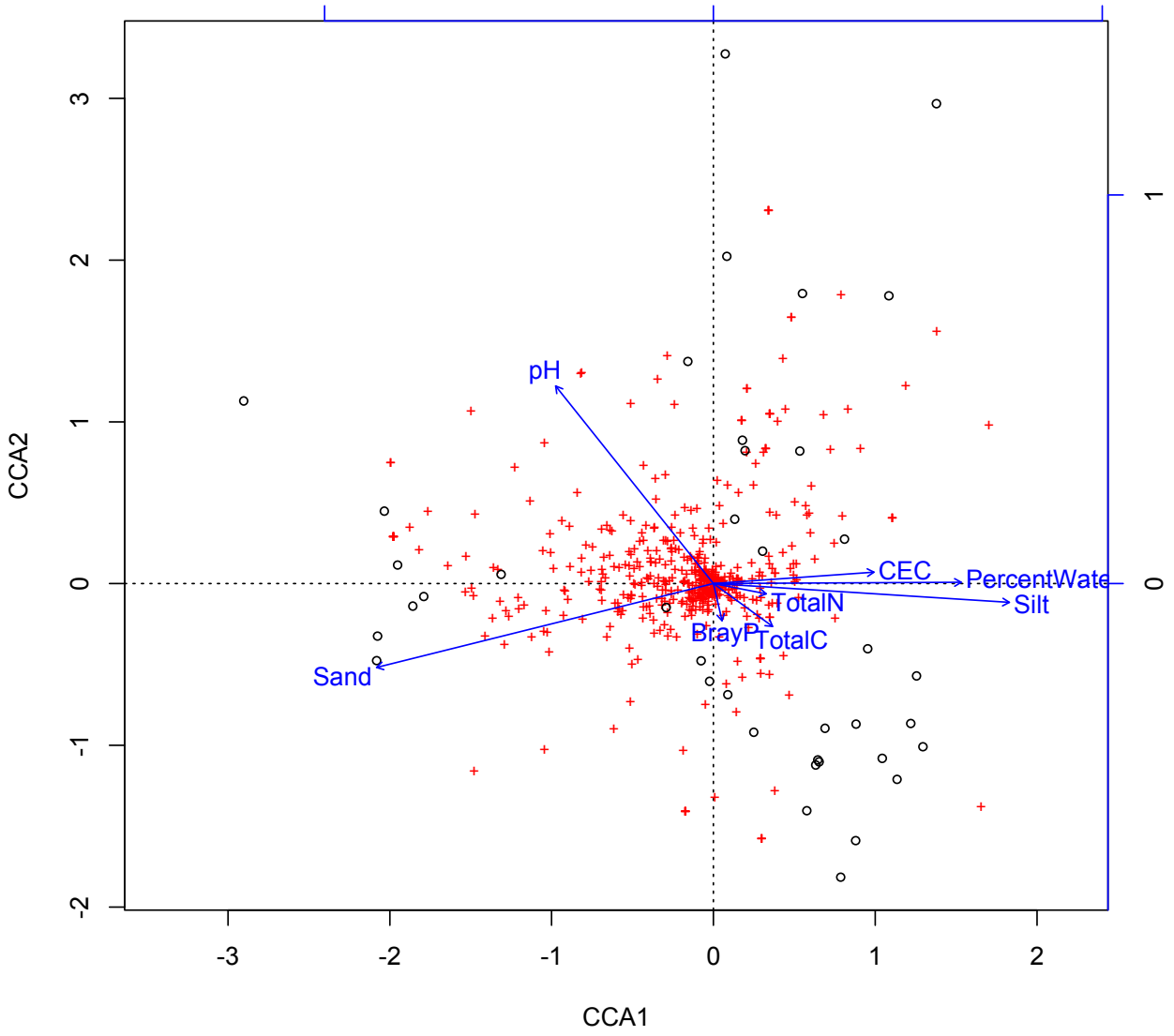


Figure S17. CCA ordination with environmental variables of the GeoChip phosphorus subset dataset.

CHAPTER 4: Investigating diatom diversification dynamics

Abstract

Diatoms are the most diverse group of marine phytoplankton. They carry out crucial ecosystem functions, including about one fifth of total global photosynthesis. This study improves our understanding of the processes that led to diatoms' extreme diversity and represents the first application of diversification modeling to a large microbial group, as past diversification studies have focused on plant and animal groups. We utilized publically available diatom sequences to build a diatom phylogenetic tree and then grafted environmental marine diatom sequences from the global Tara Oceans study onto the phylogeny. We combined this tree with a Bayesian model to estimate diatom species numbers and Modeling Evolutionary Diversification Using Stepwise AIC (MEDUSA) to look for significant increases or decreases in diversification within the four currently accepted morphological types of diatoms: radial centric, polar centric, araphid pennate, and raphid pennate. Our modeling agrees with recent molecular-based studies and shows that these four diatom morphological types are paraphyletic. We also identified several shifts in net diversification rates within the diatom phylogenetic tree, and we discuss how these rate increases correspond to the evolution of the different diatom morphologies. Recent studies have focused on utilizing diatoms to produce next generation biofuels as a feasible alternative to both fossil fuels and terrestrial-based biofuel feedstocks. Future directions for the genetic manipulation of diatoms to increase their biofuel yield have also been proposed. These proposals have led to an increased interest in diatom genetic diversity and the desire to better understand the diversification dynamics that led to their extreme diversity.

Introduction

Investigating diversification dynamics is key to understanding how biodiversity is distributed globally and how ecological communities are formed. Exploring why speciation and extinction rates of organisms vary over time, space, and different types of organisms aims to improve our understanding of how biological diversity is generated (Morlon 2014). Diversification is a predictor of species abundance distributions, species-area relationships, and distance-decay relationships (Morlon et al. 2014). Diversification trajectories can be inferred from molecular phylogenies by utilizing modeling approaches that have recently been proposed (reviewed in Morlon 2014). This study represents the first application of diversification modeling to a large microbial group, as past diversification studies have focused on plant and animal groups (e.g., Arakaki et al. 2011, Jetz et al. 2012).

Diatoms (Bacillariophycea within the division Heterokontophyta, also known as Stramenopiles) are unicellular algae that inhabit many habitats, marine, freshwater, and terrestrial. Diatoms are very diverse, possibly with as many as 100,000 total species (Falciatore and Bowler 2002), and they perform a wide variety of ecosystem functions, such as playing key roles in biogeochemical cycling. Marine diatoms, the focus of this study, occur in the ocean as deep as photosynthetically available radiation is able to penetrate (Falciatore and Bowler 2002).

In addition to the crucial ecosystem functions carried out by diatoms, diatoms have recently attracted attention from researchers who hope to utilize them to produce next generation

biofuels. Harnessing diatoms as an alternative to fossil fuels and to land-based biofuel feedstocks could potentially avoid many of the negative environmental and agricultural consequences of fossil fuels and land-based biofuel feedstocks. Diatoms would be able to replace fossil fuels using less than 5% of the land area of the United States (Levitan et al. 2014) and have already been shown to be productive in mass cultures (Dismukes et al. 2008). Compared to terrestrial biofuel feedstocks, diatoms are more efficient at collecting sunlight, can be converted to liquid fuels using simpler technology, and offer more secondary uses (Dismukes et al. 2008).

Future directions for the genetic manipulation of diatoms and the improvement of biotechnological pathways to increase their biofuels yield have been proposed (Levitan et al. 2014). These proposals have led to an increased interest in the study of diatom genetic diversity and to interest in better understanding the diversification dynamics that led to their extreme diversity.

Diatom Ecology

Diatoms are characterized by their external wall composed of amorphous silica [$(SiO_2)n(H_2O)$]. This cell wall, known as a frustule, is constructed of two halves, with the smaller half fitting into the larger half (Falciatore and Bowler 2002). Unlike smaller nano- or pico-plankton that have higher surface to volume ratios and more efficient exploitation of nutrients, diatoms have low surface to volume ratios. This means they need nutrient-rich conditions to grow and often dominate phytoplankton communities in such conditions (Sarthou et al. 2005).

Diatoms are found in a diverse variety of shapes and vary by three orders of magnitude in size (similar to land plants) (Falciatore and Bowler 2002). Small-celled diatom species (5-50 um) are most abundant when nutrients are abundant and light intensity is optimal for photosynthesis. This is typically in the beginning of spring and autumn. When these optimal conditions change, these small-celled species fall out of the photic zone (Falciatore and Bowler 2002). In contrast, giant diatoms (up to 2-5 mm) are ubiquitous in oceans and their populations show less variability across seasons. The silica cell walls of these giant diatoms are a major component of ocean floor sediment, and they play key roles in ocean biogeochemistry (Falciatore and Bowler 2002).

Role in Biogeochemical Cycles

Diatoms carry out about 20 percent of total photosynthesis on earth (Nelson et al. 1995, Armbrust 2009), and about 40 percent of the primary production in the world's oceans (Sarthou et al. 2005). Each year oceanic diatom photosynthesis produces as much organic carbon as all rainforests combined (Nelson et al. 1995, Field et al. 1998, Armbrust 2009). In contrast to the carbon fixation carried out by trees, the organic carbon generated by diatoms serves as the base of marine food webs (Sarthou et al. 2005). Diatoms are responsible for large vertical movement of primary production from the upper ocean to the deep ocean, during major seasonal phytoplankton blooms (Sarthou et al. 2005). Diatoms are also major players in the biogeochemical cycles of other macro-nutrients, including nitrogen, phosphorus, and iron. They also play a key role in the cycling of silicon, due to their growth requirements for silicon, due to their silica cell walls (Sarthou et al. 2005).

Morphological Classifications

Diatoms have been split into two major groups based on the morphology of their frustules. The first group, centric diatoms, have radially symmetrical frustules. In contrast, the

second group, pennate diatoms, are elongated and have bilaterally symmetrical frustules (Falciatore and Bowler 2002).

These two major groups are then further split into four morphological groups, based on structural details of their valves and girdle bands (Kooistra et al. 2009): radial centric, polar centric, araphid pennate, and raphid pennate diatoms. These four diatom types are currently accepted as the formal classification of diatoms in the literature, based on numerous morphological and molecular studies (Fritsch 1935, Simonsen 1979, Sims et al. 2006, Kooistra et al. 2009, Theriot et al. 2010). However, recent molecular-based studies of diatom phylogenies have revealed that these diatom types are not strictly monophyletic (Kooistra et al. 2009, Theriot et al. 2010, see "Evolutionary History of Diatoms" section below).

Radial centric diatoms have circular valves, striae that radiate from their centers, labiate processes are along their valve margin, and they are generally planktonic. Polar centric diatoms have polar valves, striae that radiate from their centers, labiate processes in their valve margin or center, and they are also generally planktonic. Araphid pennate have elongate valves, parallel striae, labiate processes and pore fields are near their valve apices, and they are typically epiphytic. Lastly, raphid pennate diatoms have elongate valves and parallel striae.

Evolutionary History of Diatoms

Diatoms have a large presence in the fossil record due to their silica cell walls, and they have numerous morphological features that were the basis for character-based systematics studies long before molecular phylogenetic studies were possible (Sims et al. 2006). The oldest diatom fossils are from the Jurassic, and well-preserved fossils are available from the Lower Cretaceous. These oldest fossils are identifiable as centric diatoms, and some may be linked to extant radial centrics and polar centrics. Pennate diatom fossils are found in the late Cretaceous, and raphid diatoms are found in the Paleocene (Sims et al. 2006).

Based on the diatom fossil record, Fritsch (1935) and Simonsen (1979) hypothesized that pennate diatoms evolved from centric diatoms, because they were found later in the fossil record. Simonsen also observed from the fossil record that centric diatoms appeared to be paraphyletic (Sims et al. 2006). Both of these observations were confirmed by the first molecular-based study of the diatom phylogeny (Medlin et al. 1993).

More recent phylogenies built from diatom SSU rRNA genes show that radial centrics are the most ancestral type of diatoms and are most likely paraphyletic. Radial centrics then gave rise to polar centrics, which are also paraphyletic. Pennates evolved from one of the lineages of polar centrics. Within pennates, araphid pennates are most likely paraphyletic and raphid pennates are monophyletic (Kooistra et al. 2009, Theriot et al. 2010). In addition, an epiphytic lifestyle is ancestral for pennates, with a planktonic lifestyle acquired four times independently and active locomotion acquired once in raphid pennates (Kooistra et al. 2009).

Despite the large number of diatom fossils in the fossil record as well as the recent studies that have focused on inferring diatom phylogenies from molecular data, further molecular data are needed to infer a more complete diatom phylogeny. For instance, despite sequencing the small ribosomal subunit gene, the large subunit of the ribulose-bisphosphate carboxylase gene, and the photosystem II gene of 136 different diatoms, there remained ambiguities in Theriot's et al. (2010) phylogenies. In that study, the chloroplast data weakly supported the monophyly of polar centrics, but the small ribosomal subunit data weakly rejected this monophyly (Theriot et al. 2010). The data also showed that there might be an unrecognized clade of araphid pennates sister to the remaining pennates in their phylogenies. Theriot et al. (2010) point out that the

biggest obstacle to inferring an accurate diatom phylogeny has been the fact that the parts of the diatom tree that are the least resolved correspond to the most under-sampled types of diatoms.

We aim to make the phylogeny more complete and more suitable for diversification modeling by grafting environmental marine diatom sequences onto a diatom phylogeny that is not too poorly resolved. While it is important to be aware of these ambiguities in the current diatom phylogeny, improving these ambiguities in the overall diatom phylogeny is not the goal of this study.

Objectives

We sought to study diatom evolutionary history by: (1) building a diatom phylogenetic tree with more environmental diatom sequences than previously published phylogenies, (2) making the diatom phylogeny more complete by grafting in marine diatom environmental sequences of the variable V9 domain of the mitochondrial small-subunit (SSU) rDNA, in order to enable marine diatom diversification modeling on a more complete phylogeny, and (3) identifying major shifts in lineage diversification rates (significant increases or decreases in speciation and extinction rates) during the evolution of marine diatoms.

Methods

Tara Oceans Dataset

In order to carry out these objectives, we utilized the unique Tara Oceans dataset. The Tara Oceans expedition was an unprecedented effort that collected 35,000 samples from 210 ocean sampling sites containing millions of marine organisms. Several high-impact studies have already been published based on the collected data. For instance, quantitative double-stranded viral-fraction metagenomes and whole viral community morphological data sets from the Tara Oceans dataset were utilized to assess structures of viral marine communities by Brum (2015). The authors established a global ocean viromic dataset and found that viral communities were passively transported by ocean currents and locally structured by environmental conditions that affect host community structure. Another study of the Tara Oceans data found that vertical stratification of the photic zone microbiome was mostly driven by temperature, not environmental factors or geography (Sunagawa et al. 2015). Similarly, Lima-Mendez et al. (2015) studied the global plankton photic zone interactome and found that environmental factors are incomplete predictors of community structure. Instead, plankton functional types and phylogenetic groups were nonrandomly distributed and driven by local and global patterns (Lima-Mendez et al. 2015).

Within the large Tara Oceans collection effort, microscopic plankton were sampled methodically at 210 sites at up to 2000 m in depth in all major ocean regions from 2009 to 2013 (Bork et al. 2015). Sampling was carried out in the water column focused on the upper layer of the ocean enriched by sunlight (surface down to 200 m), although the deeper twilight layer was also sampled (down to 2000 m) (Bork et al. 2015, Pesant et al. 2015). The samples were collected on board a research schooner and then subjected to systematic data processing on land (Bork et al. 2015). After data processing this collection effort resulted in 63,371 diatoms barcoding sequences of the variable V9 domain of the mitochondrial SSU rDNA. These marine diatom sequences are biased toward marine diatoms present in the water column as opposed to benthic diatoms, due to the sampling design of the Tara Oceans project. We combined these Tara

Oceans diatom V9 sequences with published diatom databases to investigate marine diatom diversification dynamics.

Swarm Clustering

The 63,371 diatom V9 sequences from the Tara Oceans dataset were clustered using the swarm approach described by Mahe et al. (2014). This approach seeks to avoid the two major flaws of typical clustering method: designating arbitrary clustering thresholds and input-order dependency caused by centroid selection. Swarm clustering addresses these issues by clustering similar amplicons using a local threshold and then by using each cluster's internal structure and amplicon abundances to refine the results (Mahe et al. 2014). This clustering method yielded 3,875 biologically meaningful operational taxonomic units (OTUs) from the Tara Oceans data, with each OTU represented by the most abundant ribotype in the cluster. A data table of the global relative read abundance of each swarm was also created in the swarm clustering process. To create this abundance table, the number of sequence reads corresponding to each newly defined swarm in the individual samples from each location were added together to arrive at the total global abundance of reads corresponding to each swarm.

Inferring the Phylogenetic Tree

In order to carry out the diversification modeling, the main objective of this study, a diatom phylogenetic tree was needed. This phylogenetic tree needed to include V9 sequences that were able to be aligned with the environmental marine diatom V9 sequences from the Tara Oceans dataset, in order to make the phylogeny more complete and thus study marine diatom diversification. However, the published diatom phylogenies available at the time of this study (Theriot et al. 2009, Theriot et al. 2010) did not include the necessary V9 segment.

Therefore, while the aim of the study was not to better resolve the existing overall diatom phylogenetic tree, but rather to make the phylogeny more complete and more suitable for diversification modeling by grafting environmental marine diatom sequences onto a diatom phylogeny that is not too poorly resolved, we still needed to infer a new phylogenetic tree. A phylogenetic tree of global diatom diversity was built from the 3,621 diatom sequences available in the October 2014 release of the Protist Ribosomal Reference Database (PR2) based on GenBank 203 (Guiollou et al. 2013, Chevenet et al. 2006, Chevenet et al. 2010), as well as the 1,776 additional environmental diatom SSU sequences available in GenBank at the time of the study (December 16, 2014) that were not listed in PR2. These sequences were aligned with PyNAST (Caporaso et al. 2010).

The phylogenetic tree was inferred using FastTree with the following options (./FastTree -nt -gtr -cat 4) (Price et al. 2009). FastTree infers phylogenetic trees from nucleotide sequences using approximately maximum-likelihood methods. FastTree is an appropriate algorithm to use for microbial datasets because it can handle alignments with up to a million sequences without necessitating exorbitant amounts of time and memory. It is also much more accurate than distance-matrix methods that have otherwise been used for large alignments (Price et al. 2009). The settings that we stipulated utilized a generalized time-reversible (GTR) model of nucleotide evolution. Additionally, the algorithm used a single rate of evolution for each site (the CAT approximation), in order to account for varying rates of evolution across sites.

The phylogenetic tree was converted to an ultrametric tree, a requirement of the diversification modeling, using the chronopl function in the Ape package of R (Paradis et al. 2004, Harmon et al. 2008). The chronopl function works by estimating the node ages of

phylogenetic trees utilizing a semi-parametric method based on penalized likelihood (Sanderson 2002). The default of an age of 1 was assumed for the root, and the ages of the other nodes on the tree were estimated relative to the root (Paradis et al. 2004). Therefore, while the tree was not dated relative to the fossil record and the diversification rates are thus not given in units of time, the branch lengths of the phylogenetic tree are interpretable as mean numbers of substitutions per site.

Placement of Environmental Sequences

The aim of the study was not to better resolve the existing overall diatom phylogenetic tree, but rather to make the phylogeny more. The diversification modeling analyses require a more complete phylogeny and estimates of overall global diversity for the taxa of interest. Therefore, the environmental marine diatom V9 sequences from the Tara Oceans dataset were not included in the original inferring of the phylogenetic tree, but rather grafted onto the existing phylogeny.

In order to place the Tara Oceans environmental diatom sequences in the 3,875 swarm OTUs onto the global diatom phylogenetic tree, the Tara OTUs were aligned with the PR2 and GenBank sequences using PyNAST and then placed onto the global diatom tree within the constraints of the preexisting tree topology using FastTree's constrained topology search.

Diversification Modeling

We then used the Modeling Evolutionary Diversification Using Stepwise AIC (MEDUSA) in version 2.0.6 of the Geiger R package (Alfaro et al. 2009, Harmon et al. 2008), which detects diversification rate shifts from phylogenetic data. One advantage of MEDUSA is that it does not require species-level phylogenies. Instead, it requires clade-level phylogenies along with the species richness of the various unresolved clades.

Diversification modeling using MEDUSA requires richness information be provided as an input, if the tree is not completely sampled (Alfaro et al. 2009). The richness input in MEDUSA links species richness with lineages within the tree. Therefore, if a large clade of many known species is represented within the tree by a single tip, the total diversity of the clade is included in the richness input. Due to the fact that the diatom tree is not completely sampled, we estimated overall diatom diversity as well as the diversity of specific clades within the tree. In order to calculate these estimates, we used the relative abundance of each of the Tara diatom V9 sequences in the clade (from the global abundance table described above) and the Bayesian Diversity Estimation Software described by Quince et al. (2008) using the Metro Log Normal abundance distribution. The Bayesian Diversity Estimation algorithm fits an abundance distribution to the observed taxa and their abundances in order to infer the number of taxa in the overall community. This approach was developed in order to better estimate extant microbial diversity, given that even metagenomic sequencing can only sample a small portion of an environmental microbial community.

We defined three different sets of clades in order to assess the effect of clade choice on the estimation of significant shifts in diversification rates. In order to select the three sets of clades, we identified all of the nodes in the phylogenetic tree by node number and then created a table listing all of the nodes, the number of terminal clades descendent from each node, and the percentage of these terminal clades that were composed of Tara sequences. We then varied the cutoffs of the two criteria in the table (the number of terminal clades descendent from each node and the percentage of each clade that was composed of Tara sequences) that qualified a node for

selection and selected the nodes that fell within a specified range. We varied the number of terminal clades descendent from each node from 19 to 500 terminal clades. Nodes with a greater number of terminal clades were more basal within the tree, while nodes with fewer terminal clades were closer to the tip of the tree. We varied the percentage of each clade that was composed of Tara sequences from 10% to 90%, in order to select clades that were well represented by the environmental Tara Oceans sequences.

The chosen sets of nodes resulted in phylogenetic trees with 25, 73, and 99 terminal clades. These three sets of nodes were selected from the phylogenetic tree based on the following criteria: 25 terminal clades (between 19 and 100 terminal clades descendant from each node, descendants of each node were composed of between 25% and 70% Tara sequences), 73 terminal clades (between 19 and 500 terminal clades descendant from each node, descendants of each nodes were composed of between 30% and 90% Tara sequences), and 99 terminal clades (between 19 and 500 terminal clades descendant from each node, descendants of each node were composed of between 10% and 90% Tara sequences).

The sequences within each terminal clade were labeled as belonging to one of the four major diatom groups (polar centric, radial centric, raphid pennate, or araphid pennate), based on lineage information that was included in the original Tara Oceans dataset. This lineage information was not available for all of the Tara sequences, so sequences without lineage information were labeled as belonging to an "unknown" diatom group (also referred to as "UK"). Based on these lineage classifications, the diatom groups that the descendants of each estimated rate diversification shift belonged to were used as a first attempt at an approximation of ancestral characters at each shift. While this was not a very rigorous estimate of ancestral states, these lineage classifications were the only morphology-related data associated with the diatom sequences there were available at the time of this study.

Results

Diversification Modeling

A diatom phylogenetic tree with 8,530 tips and 8,353 internal nodes was inferred. The estimate of total global diatom diversity based on the Bayesian Diversity Estimation Software was 55,000 diatom species. In each of the 25, 73, and 99 terminal clades trees, each terminal clade is labeled to indicate whether it represents 1-49 estimated diatom species, 50-99 estimated diatom species, 100-499 diatom species, or 500-2000 diatom species (see Figures 1-3).

The MEDUSA analysis on the diatom tree with 25 terminal clades estimated two significant changes in the tempo of diversification in diatom history (Table 1, Figure 1). Both shifts correspond to increases in net diversification when compared to the background rate of diversification (r = 1.41055 [Background], 5.14068 [Shift 1], 8.11122 [Shift 2]). MEDUSA chose a Yule model of diversification for the background diversification and for both shifts.

The MEDUSA analysis on the diatom tree with 73 terminal clades estimated four significant changes in the tempo of diversification in diatom history (Table 2, Figure 2). Shifts 1 and 4 correspond to increases in net diversification, while Shifts 2 and 3 correspond to decreases in net diversification (r = 3.33899 [Background], r = 32.2825 [Shift 1], r = 2.51938 [Shift 2], r = 0 [Shift 3], r = 31.0746 [Shift 4]). MEDUSA chose a birth-death model for the background diversification and Shift 2 and a Yule model for Shifts 1, 3, and 4.

The MEDUSA analysis on the diatom tree with 99 terminal clades estimated six significant changes in the tempo of diversification in diatom history (Tale 3, Figure 3). Shifts 1, 3, and 5 correspond to increases in net diversification, while Shifts 2, 4, and 6 correspond to decreases in net diversification (r = 2.70153 [Background], 131.47900 [Shift 1], 5.76073 [Shift 2], 740.40500 [Shift 3], 22.07880 [Shift 4], 429.23100 [Shift 5], 7.30017 [Shift 6]). MEDUSA chose a Yule model for the background diversification and Shifts 3 and 5 and a birth-death model for Shifts 1, 2, 4, and 6.

Tree Comparisons

Comparing the three different trees with different numbers of analyzed terminal clades (25, 73, and 99 terminal clades), the different selection of clades resulted in slightly different tree topologies, and in two, four, and six rate shifts, respectively (Figs. 1-3). However, several of the major estimated shifts appear in more than one of the trees. For example, Shift 1 in the 73 terminal clades tree corresponds to Shift 1 in the 99 terminal clades tree, and Shift 4 in the 73 terminal clades tree corresponds to Shift 3 in the 99 terminal clades tree. Shift 2 in the 25 terminal clades tree corresponds to Shift 4 in the 99 terminal clades tree. Shift 1 in the 25 terminal clades tree also appears to correspond to Shift 2 in the 99 terminal clades tree. However, there does not appear to be an equivalent shift in the 73 terminal clades tree, and the location of this shift instead falls under the background rate in the 73 terminal clades tree. Finally, Shift 3 in the 73 terminal clades tree does not appear to correspond to any shifts in either of the other two trees.

Diatom Group Classifications

For all three of the trees, we calculated the number of Tara swarm sequences from each node and which of the four diatom groups they were classified into. We noted which diatom group the majority of the Tara sequences belonged to and what percentage of the sequences that group represented. If the majority of the sequences descending from a particular node were classified as an unknown group, then that terminal clade was labeled “UK”, and the diatom group with the second most number of Tara sequences in it was noted (Figures 1, 2, and 3). Classifying terminal clades by the diatom groups they best represent better allowed us to compare the estimated shifts in each of the three trees and to note which diatom groups were well represented in rate shifts with estimated positive or negative net diversification rates.

Shift 1 in the 73 terminal clades tree and Shift 1 in the 99 terminal clades tree are composed largely of radial centric sequences, while Shift 4 in the 73 terminal clades tree and Shift 3 in the 99 terminal clades tree are composed largely of polar centric sequences. Shift 2 in the 25 terminal clades tree and Shift 4 in the 99 terminal clades tree are composed largely of raphid pennate sequences.

Unexpectedly, classifying the terminal clades by diatom group classification majorities reveals multiple groupings of each of the four major diatom groups throughout the phylogenetic trees. That is, instead of showing one origination of polar centric diatoms, the diatom group classification on the three trees show multiple places where the majority of the Tara sequences descendent from a particular node were classified as polar centric.

Discussion

We inferred a global diatom phylogenetic trees with 8,530 tips, defined three different sets of clades in order to assess the effect of clade choice on the estimation of significant shifts in diversification rates, and estimated the overall global species richness of each of these clades. The estimate of total global diatom diversity based on the Bayesian Diversity Estimation Software was 55,000 diatom species, which is within the range of published estimates of diatom diversity of 20,000 to 100,000 estimated total species (Leblanc et al. 2012). We then performed MEDUSA analyses on the trees in order to model diatom diversification over time. For the terminal clades in three of the trees, we also noted which diatom group the majority of the Tara sequences belonged to and what percentage of the sequences that group represented.

Net Diversification Rates

Approximately half of the significant rate shifts estimated by MEDUSA across all three trees corresponded to increases in diversification rates and half corresponded to decreases in net diversification rates. Of the major estimated shifts that were conserved among multiple trees, Shift 1 in the 73 terminal clades tree and Shift 1 in the 99 terminal clades tree were composed largely of radial centric sequences, while Shift 4 in the 73 terminal clades tree and Shift 3 in the 99 terminal clades tree were composed largely of polar centric sequences. Shift 2 in the 25 terminal clades tree and Shift 4 in the 99 terminal clades tree were composed largely of raphid pennate sequences.

Comparing our results to those of other MEDUSA-based diversification studies shows similar increases in net diversification in other major groups. Alfaro et al. (2009) developed the MEDUSA approach in order to study the diversification of 44 clades of jawed vertebrates. Like our study, they identified major diversification rate increases in the evolution of jawed vertebrates. These net increases in diversification corresponded to well-known radiations, including those of modern birds, lizards and snakes, ostariophysan fishes, and eutherian mammals. Similarly, Near et al. (2013) used MEDUSA to study the evolution of spiny-rayed fishes, which account for about one-third of extant vertebrates. Clade-specific analyses revealed multiple rapid radiations, including tunas, gobies, blennies, snailfishes, and Afro-American cichilds. These radiations were not associated with a specific habitat type.

Similar to our study, Alfaro et al. also detected three significant decreases in the rate of diversification and found that diversification rates varied greatly among lineages, and Near et al. found a global decrease in lineage diversification, which corresponded to a period or morphological disparity among fossils. Alfaro et al. did also find sections of the phylogenetic tree that had nearly equal rates of speciation and extinction, which spoke to importance of faunal turnover in shaping biodiversity.

Jetz et al. (2012) built the first dated phylogeny of 9,993 extant bird species and then utilized MEDUSA to explore avian diversification. They found birds experienced a strong increase in diversification rate through time. They also carried out geographic analyses and found major differences in diversification rates between hemispheres. They hypothesized that the increasing diversification rate found in their study may have been due to the novel global scale of their analyses, with smaller clades being more geographically and ecologically bounded. Arakaki et al. (2011) used the MEDUSA approach to study cacti diversification. They found major cactus radiations occurred simultaneously as several global succulent lineages on multiple continents.

While we did not conduct location-based diversification analyses, we did work with a global dataset that allowed us to carry out analyses at the same scale as Jetz et al. and Arakaki et al.

Diversification of Diatom Groups

A few of the increases in net diversification were estimated at clades with large percentages of radial centrics (Shift 1 in the 73 terminal clades tree, Shift 1 in the 99 terminal clades tree), which are the most ancestral morphology of diatoms. These rate shifts may correspond with two major radiations of radial centrics, since radial centrics are divided into two main groups: basal radial centrics and the remaining radial centrics. Basal radial centrics are too heavy to live a planktonic lifestyle and are found in shallow coastal regions where they thrive in turbulent environments. The rest of the more descendent radial centrics experienced a rapid adaptive radiation in both freshwater and marine planktonic habitats and epiphytic communities (Kooistra et al. 2007).

Basal radial centrics gave rise to polar centrics, which are also paraphyletic and generally planktonic. The several rate shifts that were estimated at clades with large percentages of polar centrics (Shift 4 in the 73 terminal clades tree, Shifts 3 and 5 in the 99 terminal clades tree) show important accelerations with the evolution of the largely planktonic polar centrics. These separate rate shifts likely correspond to periods of rapid adaptive radiation during which several lineages developed a successful planktonic lifestyle, compared to deeper lineages that are epiphytic (Kooistra et al. 2007).

Pennates evolved from one of the lineages of polar centrics. Within pennates, araphid pennates are most likely paraphyletic and raphid pennates are monophyletic (Kooistra et al. 2009). Araphid pennates have an elongate shape, a midrib, striae perpendicular to it, apical labiate processes, and apical pore fields (Kooistra et al. 2007). However, none of these traits are uniquely shared among them. In contrast, the rate shifts that were estimated at a clade with a large percentage of raphid pennates (e.g., Shift 2 in the 25 terminal clades tree) may correspond to the acceleration that resulted in the evolution of a new diatom lifestyle. Raphid pennates are monophyletic, having acquired raphe during one event (Hasle 1974). The acquisition of this novel mode of locomotion was hugely successful, allowing raphid pennates to actively move in search of mates and nutrition and to move away from threats and too much light (Bates and Davidovich 2002, Kooistra et al. 2007). Raphid pennates are now the largest group of extant diatoms (most genera and species), even though they are the youngest lineage (Kooistra et al. 2007).

Robustness of Results

In order to study diatom diversification, we collapsed some of the clades of the diatom phylogenetic tree inferred for this study. Due to the fact that the criteria and cutoffs for selecting these clades were semi subjective (based on varying clade sizes and percentages of the clades composed of Tara Oceans environmental sequences), three different sets of clades (25, 73, and 99 clades) were selected in order to compare the effects of clade choice on the estimation of significant shifts in diversification rates. The MEDUSA modeling of rate shifts showed that despite the different numbers of selected clades, several of the estimated major appeared in multiple trees. The conservation of these rate shifts across the three different trees, despite the different clades that were selected, shows that the signal for these significant rate shifts in the diatom phylogeny is strong. The modeling of the rate shifts was not dependent on or strongly affected by the criteria for clade selection.

Additionally, the MEDUSA approach we utilized to model diversification in this study has very recently come into question. May and Moore (2016) found that the statistical behavior of MEDUSA was previously unknown, and they performed an investigation into the behavior and biases of MEDUSA as a modeling approach. They found that MEDUSA has a high false-discovery rate and provides biased estimates of diversification rate parameters. They theorized that this is caused in part by the fact that the likelihood functional in MEDUSA is incorrect and by the fact there is little to guide the specification of an appropriate AIC critical threshold for selecting one model over another. These findings cast doubt on the conclusions of many of the diversification studies that have utilized the MEDUSA modeling approach (e.g., Alfaro 2009, Arakaki et al. 2011, Jetz et al. 2012, Near et al. 2013).

Unexpectedly, classifying the terminal clades by majority morphological type revealed multiple groupings of each of the four major morphologies throughout the phylogenetic trees. For instance, instead of showing one origination of raphid pennate diatoms, the diatom group classification on the three trees show multiple places where the majority of the Tara sequences descendent from a particular node were classified as polar centric. Due to what we know from past diatom taxonomic studies (Kooistra et al. 2007, Theriot et al. 2009, Theriot et al. 2010), this may be a flawed incorrect representation of the evolution of the four major diatom groups. While Kooistra et al. and Theriot et al. have shown in their past molecular-based phylogenetic studies that the four major diatom types are paraphyletic, they have not found so many paraphylies throughout the tree. This may be a result of a less accurate or poorly resolved phylogenetic tree, possibly due to the approximately maximum-likelihood FastTree method used to build the tree for our study. Alternatively, this could be an artifact of the way that the Tara sequences were classified into diatom groups. When the dataset was collated and clustered into swarms by the Tara Oceans team of researchers, lineage information was included with each of the Tara sequences. However, because many of the Tara sequences are novel environmental sequences collected from throughout the world's oceans, lineage information was not available for all of the sequences at the time of this study. Sequences without lineage information were labeled as belonging to an "unknown" diatom group. The robustness of this diatom group classification and approximation of ancestral characters was thus reduced because of how many of the Tara sequences in each clade had available lineage information and how many had unknown lineages. For instance, in the 99 terminal clades tree, the majority of the Tara sequences in 29 of the terminal clades had unknown lineages. Of those 29 clades, in 22 of the clades had greater than 70% of the Tara sequences belonged to an unknown diatom group. The diatom group classification of these lineages was thus based on less than 30% of the sequences belonging to those clades, and the second most represented group, on which the colored labeling of the diatom groups was based (Figures 1, 2, and 3) was often only based on around 15% of the total sequences in a clade. This lack of accurate lineage information likely compromised our ability to accurately describe and analyze the diversification of specific diatom groups.

Lastly, due to the sampling design of the Tara Oceans project, the environmental V9 diatom sequences used in this study were biased toward marine diatoms present in the water column as opposed to benthic marine diatoms (Bork et al. 2015, Pesant et al. 2015). This means that, while we intended to study the diversification of the global population of marine diatoms, the environmental V9 sequences we used are unlikely to be a random sample of all marine diatoms. Therefore, our analyses are likely biased toward the diversification of phytoplanktonic marine diatoms and do not include all marine diatoms with different life histories.

Future Directions

We have several ideas for the continuation of this research. First, due to the very recent critiques of the MEDUSA modeling approach, this study could be redone as is but with an alternative diversification modeling approach. Secondly, it would be worthwhile to repeat the analyses done in this study after obtaining improved lineage information regarding the major diatom groups that the Tara sequences belong to. With so many of the Tara sequences used in this study having unknown lineage classifications, it was difficult to trust the accuracy of how the estimated rate shifts related to the evolution of each diatom group.

Thirdly, it can be difficult to differentiate between increases and decreases in net diversification rates when a phylogeny only includes extant taxa (Sanmartin and Meseguer 2016). Therefore, incorporating diatom diversity from the fossil record could improve our ability to differentiate between increases and decreases in net diversification. While this would be a difficult feat for many taxon studies, diatoms are well-suited for this type of study, as they have a large presence in the fossil record (Sims et al. 2006). Future work could incorporate data from the diatom fossil record into the molecular-based phylogeny used in this study to better differentiate between increases and decreases in net diversification rates.

Lastly, following the work of Jetz et al. (2012) and Arakaki et al. (2011), this study could be continued by taking a geographic approach to analyze diatom diversification patterns. While we already analyzed the global diatom dataset, future work could investigate differences in diversification rates based on the different Tara Oceans sampling locations where the environmental diatom sequences were collected.

Conclusion

MEDUSA analyses on the global diatom phylogenetic tree we built yielded estimates of diversification rate shifts across the tree. We discussed which of these shifts appear to correspond with the radiations of major diatom groups. We specifically discuss the rate shifts within the context of the four main recognized groups of diatoms, radial centric, polar centric, and raphid pennate diatoms, which we found to be paraphyletic. We discussed several factors that may have affected the robustness of our results, including the fact that the MEDUSA approach we utilized to model diversification in this study has very recently come into question, that the lineage information we used was not available for all of the sequences at the time of this study, and that the environmental V9 diatom sequences used in this study were biased toward marine diatoms present in the water column as opposed to benthic marine diatoms. We also discussed several ideas for the continuation of this research, including repeating the study with an alternative diversification modeling approach, obtaining improved lineage information regarding the major diatom groups that the Tara sequences belong to, incorporating diatom diversity from the fossil record to better differentiate between increases and decreases in net diversification, and taking a geographic approach to analyze diatom diversification patterns. Despite the possible methodological limitations of this study and the opportunities for additional future research, this study improves our understanding of the diversification dynamics that led to diatoms' extreme diversity. This study also represents the first application of diversification modeling to a large microbial group.

Tables

Table 1. Net diversification rate shifts found in MEDUSA analysis of the 25 terminal clades tree*

Step Number	Shift Number	Shift Node	Model	r	Ln Lik	AICc
1	Background	26	Yule	1.41055	-3.596024	143.9588
2	1	28	Yule	5.14068	-34.71063	135.5706
3	2	36	Yule	8.11122	-21.91443	131.8375

Table 2. Net diversification rate shifts found in MEDUSA analysis of the 73 terminal clades tree*

Step Number	Shift Number	Shift Node	Model	r	Ln Lik	AICc
1	Background	73	bd	3.33899	-75.34278	386.1953
2	1	98	Yule	32.2825	-16.9711	373.5017
3	2	105	bd	2.51938	-60.84325	360.7239
4	3	53	Yule	0	0	355.5924
5	4	120	Yule	31.0746	-10.5456	351.4207

Table 3. Net diversification rate shifts found in MEDUSA analysis of the 99 terminal clades tree*

Step Number	Shift Number	Shift Node	Model	r	Ln Lik	AICc
1	Background	100	Yule	2.70153	-7.784727	347.7334
2	1	140	bd	131.479	-8.233126	304.5202
3	2	103	bd	5.76073	-26.2412	285.3592
4	3	170	Yule	740.405	-8.554171	274.3546
5	4	117	bd	22.0788	-12.84096	263.4836
6	5	136	Yule	429.231	2.74403	257.8313
7	6	150	bd	7.30017	-43.91171	247.0627

* Model = diversification model preferred by MEDUSA (bd = birth-death), r = net diversification rate, Ln Lik = log likelihood value, AICc = sample-size corrected Akaike information criterion threshold computed by MEDUSA

Figures

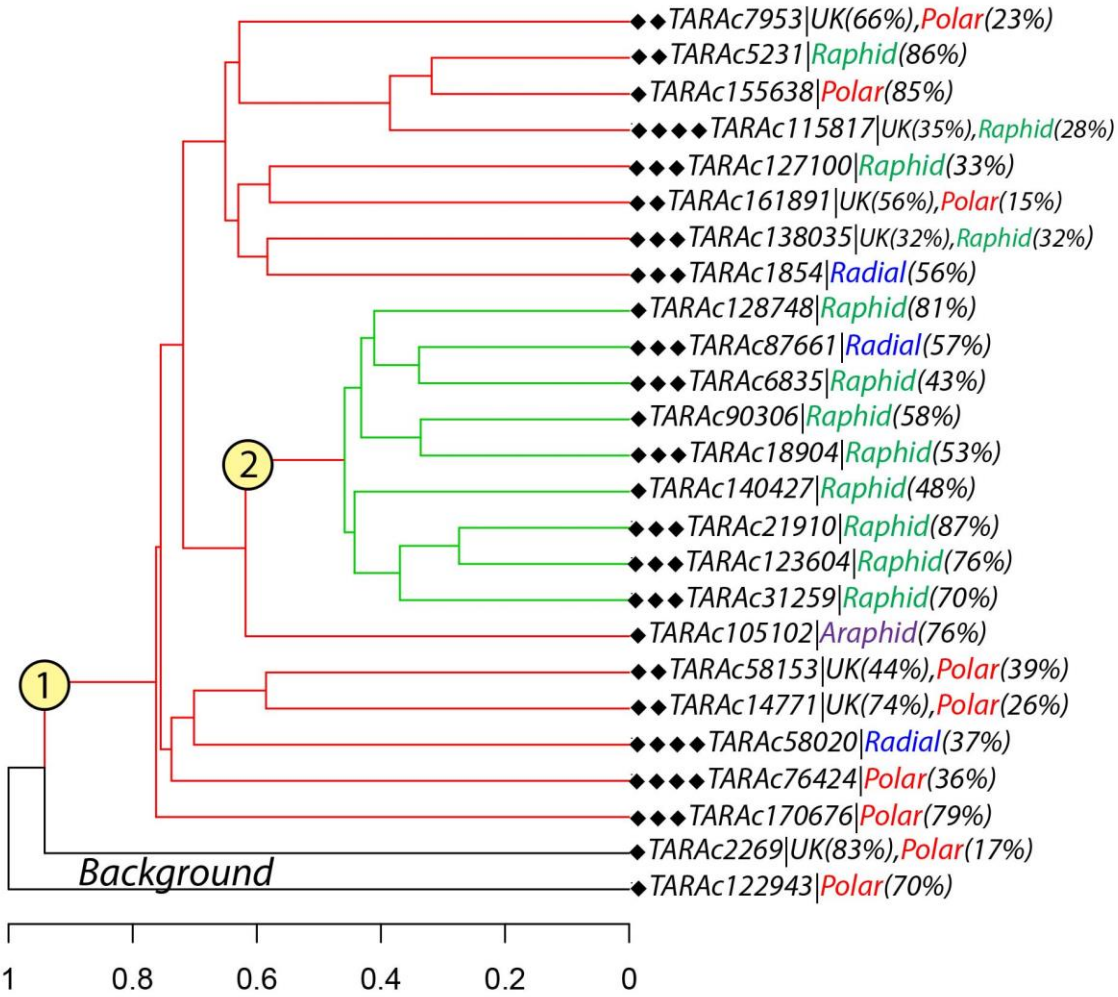


Figure 1. 25 terminal clades maximum-likelihood phylogenetic tree plotted with significant net diversification rate shifts (numbered yellow circles). Each terminal clade is labeled with a representative sequences (e.g., "TARAc73637"), as well as the diatom group that most (percentage included in the label) of the sequences in that clade belong to (e.g., Polar 70%). If the diatom group of most of the sequences in a clade is unknown (labeled as "UK"), then the second most represented diatom group is also listed. The black diamonds on each tip denote the estimated diversity of each terminal clade (◆1-49 species, ◆◆50-99 species, ◆◆◆100-499 species, ◆◆◆◆500-2000 species). For the x-axis, a default age of 1 was assumed for the root, and the ages of the other nodes on the tree were estimated relative to the root. These branch lengths are interpretable as mean numbers of substitutions per site.

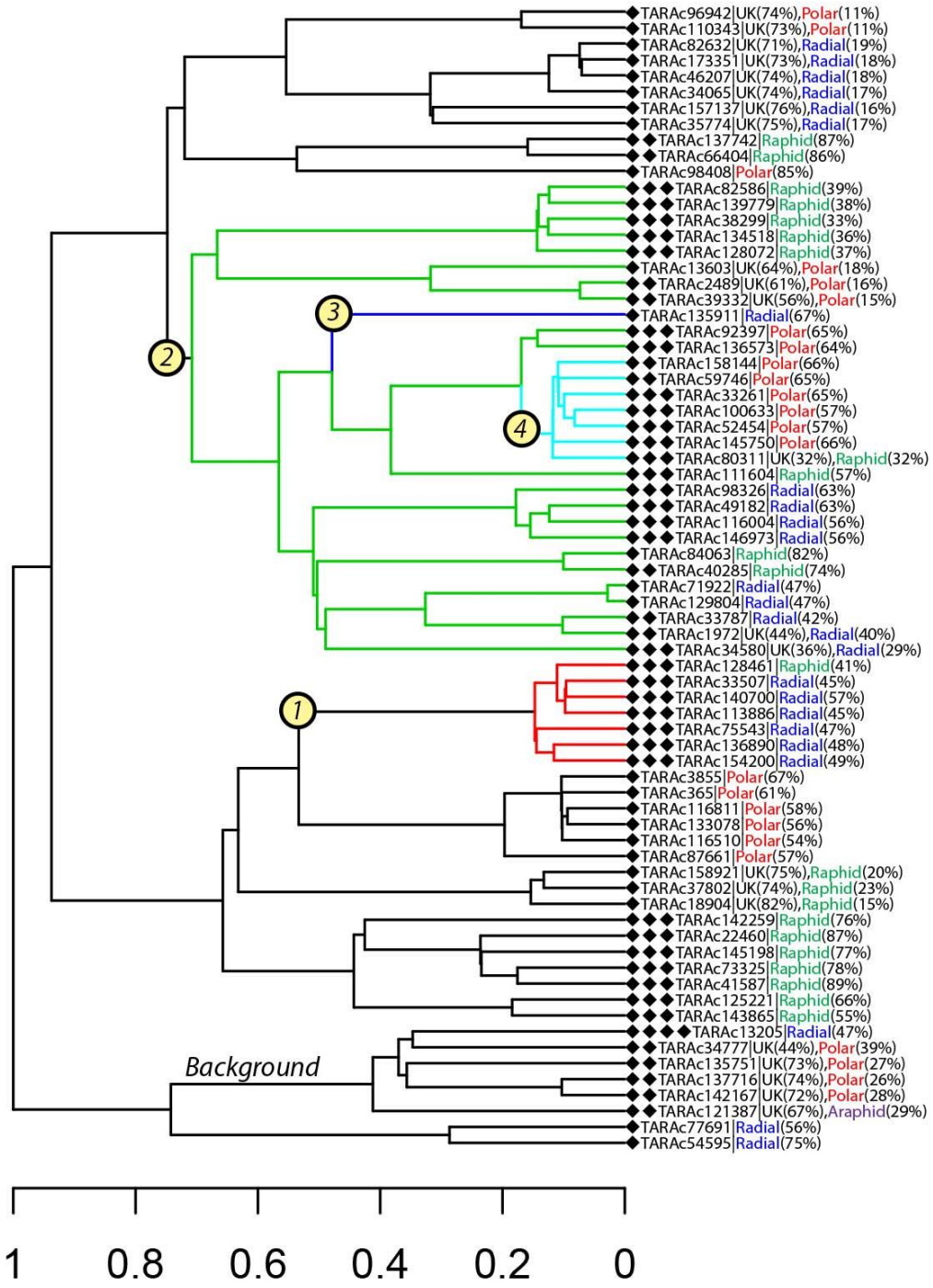


Figure 2. 73 terminal clades maximum-likelihood phylogenetic tree plotted with significant net diversification rate shifts (numbered yellow circles). Each terminal clade is labeled with a representative sequences (e.g., "TARAc73637"), as well as the diatom group that most (percentage included in the label) of the sequences in that clade belong to (e.g., Polar 70%). If the diatom group of most of the sequences in a clade is unknown (labeled as "UK"), then the second most represented diatom group is also listed. The black diamonds on each tip denote the estimated diversity of each terminal clade (◆1-49 species, ◆◆50-99 species, ◆◆◆100-499 species, ◆◆◆◆500-2000 species). For the x-axis, a default age of 1 was assumed for the root, and the ages of the other nodes on the tree were estimated relative to the root. These branch lengths are interpretable as mean numbers of substitutions per site.

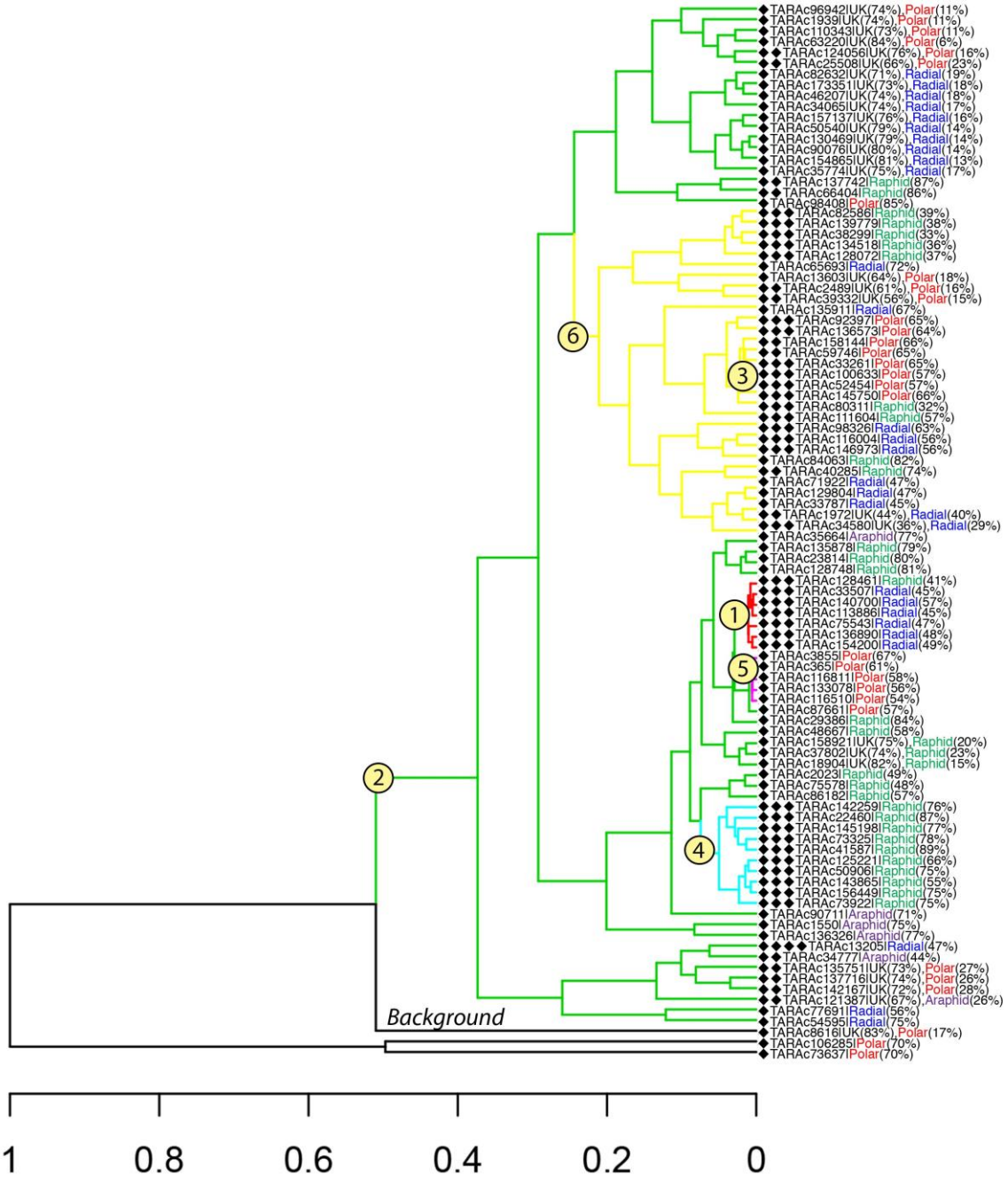


Figure 3. 99 terminal clades maximum-likelihood phylogenetic tree plotted with significant net diversification rate shifts (numbered yellow circles). Each terminal clade is labeled with a representative sequences (e.g., "TARAc73637"), as well as the diatom group that most (percentage included in the label) of the sequences in that clade belong to (e.g., Polar 70%). If the diatom group of most of the sequences in a clade is unknown (labeled as "UK"), then the second most represented diatom group is also listed. The black diamonds on each tip denote the estimated diversity of each terminal clade (◆1-49 species, ◆◆50-99 species, ◆◆◆100-499 species, ◆◆◆◆500-2000 species). For the x-axis, a default age of 1 was assumed for the root, and the ages of the other nodes on the tree were estimated relative to the root. These branch lengths are interpretable as mean numbers of substitutions per site.

CHAPTER 5. Conclusion

Microbial taxonomic, phylogenetic, and functional diversity remain understudied and poorly understood compared to our knowledge of macrofaunal diversity. This is largely due to microbes' small size, particular evolutionary history, and reproductive mechanisms, which are all very different than those of macro-organisms (Taylor et al. 1999, Eisen 2000, Hughes et al. 2001, Bowler et al. 2008, Chan et al. 2012, Jousset et al. 2013). However, advances in molecular technologies are beginning to close this gap and allowing ecologists to make great progress in studying microbial community diversity.

This dissertation utilized these new technologies to: (1) address the challenges of quantifying and comparing modern microbial data, (2) elucidate the changes to soil microbial diversity caused by widespread land-use change in tropical ecosystems, and (3) model the unique evolution of marine diatoms.

Diversity profiles, introduced in Chapter 2, allowed for the exploration of the importance of taxa rarity data and similarity data in comparing different microbial communities. Most diversity indices do not account for rarity or similarity information, or they do so in a way that is not obvious to the user. Chapter 2 illustrated that, for several of the datasets, analyzing different weightings (q values) of taxa rarity changed the conclusions made concerning the comparative diversity of the different study communities. This was also true for the inclusion of taxa similarity information. For example, for some datasets, when phylogenetic similarity was included in diversity calculations, the diversity calculation decreased. This indicates that the community contained member taxa that were closely related to each other, but this insight into the community's structure would not have been perceivable without the inclusion of similarity information. This is particularly meaningful for analyzing microbial diversity data, because the phylogenetic distribution of a microbial assemblage can influence ecosystem processes via differences in the suites of traits present.

These findings point to the need to utilize tools such as diversity profiles, which include information on taxa rarity and similarity to analyze microbial data, if we are to fully understand how community composition and phylogenetic relatedness explain ecological processes. While this dissertation analyzed both experimental and observation datasets from all microbial domains, it did not include more than one of each type of dataset. Chapter 2 particularly highlights that we would gain a clearer understanding of how well diversity profiles allow us to quantify and compare microbial data, if future work were able to compare multiple bacterial, archaeal, fungal, and viral datasets that were collected with different study designs and hypotheses. This is feasible now that technological advances allow ecologists to more easily conduct replicated multi-factor studies of microbial community diversity.

Chapter 3 discovered and quantified multiple dimensions of bacterial, archaeal, and fungal diversity in five different land-use types (Primary Forest, Secondary Forest, Oil Palm, Rubber, and Rice) throughout a dipterocarp forest landscape in Peninsular Malaysia. Analyses indicated that when Primary Forest soils were converted to alternate land-use types, some rare taxa of soil microbes were lost and unable to persist in the microbial community in these anthropogenically altered land uses. However, comparing the soil microbial communities present at the same time in different land-use types only tells us so much about how land-use change alters the original microbial community in a primary forest, due to the great variation of soil microbial communities across time and space (Green and Bohannan 2006). Therefore, it would

be valuable to conduct a longitudinal study in the future in which soil from the same area is sampled before, during, and after a primary forest is converted to an alternate land-use type. This would reduce the effects that the naturally high turnover of plant populations across space in tropical forest systems has on the soil microbial communities sampled and lead to a greater understanding of how anthropogenic disturbances affect soil microbial communities.

Our analyses of microbial data in Chapter 3, specifically the 16S and ITS datasets, revealed that geographic distance had a significant effect on 16S diversity and an even more significant effect on ITS diversity. These results are similar to other recent studies that have found that soil bacterial community similarity decreased with distance (i.e., Monroy et al. 2012). However, they are in contrast to a recent study in the Amazon that found that local soil bacterial diversity increases after conversion, but that communities become more similar across space (Rodrigues et al. 2013). The findings indicate that it would be beneficial for future research to better tease apart the relationship between soil microbial diversity in Southeast Asian forest ecosystems and geographic distance. This could be done by sampling at a greater variety of spatial scales in all of the studied land-use types.

Chapter 4 explored the diversification dynamics of marine diatoms and the significant increases and decreases in diversification rates that were estimated using a phylogeny inferred from molecular data. This chapter's findings for diatom diversification correspond to similar increases in net diversification in other major groups, such as jawed vertebrates, spiny-rayed fishes, and birds. This work has led to several ideas for the continuation of the diatom diversification research. First, due to the very recent critiques of the MEDUSA modeling approach, this study could be repeated as is but with an alternative diversification modeling approach. Secondly, it would be worthwhile to repeat the analyses done in this study after obtaining improved lineage information for the Tara sequences dataset. With so many of the Tara sequences used in this study having unknown lineage classifications, it was difficult to trust the accuracy of how the estimated rate shifts related to the evolution of each diatom group. Lastly, following the work of Jetz et al. (2012) and Arakaki et al. (2011), this study could be continued by taking a geographic approach to analyze diatom diversification patterns. While we already analyzed the global diatom dataset, future work could investigate differences in diversification rates based on the different Tara Oceans sampling locations where the environmental diatom sequences were collected.

Overall, this dissertation explored the effects of anthropogenic disturbance and environmental change on multiple dimensions of microbial biodiversity. The three somewhat disparate questions investigated addressed how microbial diversity evolves, how microbial diversity reacts to environmental change, and the methods that allow us to quantitatively study and compare the diversity of different microbial communities. In answering these questions, the dissertation provides insight into how the world's rapidly changing environment will affect microbial diversity in the future and serves as a jumping-off point for additional studies investigating the effects of global change on microbial communities. Documenting the effects of global change on microbial communities is especially relevant given the critical roles that microbial diversity plays in provisioning ecosystem services, and the fact that recent advances in technology allow us to study microbial diversity in novel ways.

REFERENCES

Abdullah MJ, Ibrahim MR. The incidence of forest fire in Peninsular Malaysia: History, root causes, prevention and control. PREVENTION AND CONTROL OF FIRE IN PEATLANDS. 2002 Mar:20.

Achard F, Eva HD, Stibig HJ, Mayaux P, Gallego J, Richards T, Malingreau JP. Determination of deforestation rates of the world's humid tropical forests. Science. 2002 Aug 9;297(5583):999-1002.

Ackerly DD, Cornwell WK: A trait-based approach to community assembly: partitioning of species trait values into within- and among-community components. Ecol Lett 2007, 10:135–145.

Alfaro ME, Santini F, Brock C, Alamillo H, Dornburg A, et al. 2009. Nine exceptional radiations plus high turnover explain species diversity in jawed vertebrates. Proceedings of the National Academy of Sciences 106:13410-13414.

Alkema W, Boekhorst J, Wels M, van Hijum SA. Microbial bioinformatics for food safety and production. Briefings in bioinformatics. 2016 Mar 1;17(2):283-92.

Allison SD, Martiny JB. Resistance, resilience, and redundancy in microbial communities. Proceedings of the National Academy of Sciences. 2008 Aug 12;105(Supplement 1):11512-9.

Amend AS, Seifert K, Samson R, Bruns TD: Indoor fungal composition is geographically patterned and more diverse in temperate zones than in the tropics. Proc Natl Acad Sci U S A 2010, 107:13748–13753.

Arakaki M, Christin PA, Nyffeler R, Lendel A, Eggli U, et al. 2011. Contemporaneous and recent radiation of the world's major succulent plant lineages. Proc Nat Acad Sci. 108(20):8379-8384.

Armbrust EV. 2009. The life of diatoms in the world's oceans. Nature. 459:185-192.

Arrigo KR. Marine microorganisms and global nutrient cycles. Nature. 2005 Sep 15;437(7057):349-55.

Azam F, Worden AZ. Microbes, molecules, and marine ecosystems. Science. 2004 Mar 12;303(5664):1622-4.

Baas-Becking LG. Geobiologie; of inleiding tot de milieukunde. WP Van Stockum & Zoon NV; 1934.

Bailey-Serres J, Fukao T, Ronald P, Ismail A, Heuer S, Mackill D. Submergence tolerant rice: SUB1's journey from landrace to modern cultivar. Rice. 2010 Sep 1;3(2-3):138-47.

Balmford A, Bruner A, Cooper P, Costanza R, Farber S, Green RE, Jenkins M, Jefferiss P, Jessamy V, Madden J, Munro K. Economic reasons for conserving wild nature. science. 2002 Aug 9;297(5583):950-3.

Bates SS and Davidovich NA. 2002. Factors affecting the sexual reproduction of diatoms, with emphasis on Pseudo-nitzschia spp. In LIFEHAB workshop: life history of microalgal species causing harmful algal blooms (Ed. by E. Garcés, A. Zingone, B. Dale, M. Montresor & B. Reguera)(pp. 31-36).

Battistuzzi FU, Hedges SB. A major clade of prokaryotes with ancient adaptations to life on land. Molecular biology and evolution. 2009 Feb 1;26(2):335-43.

Beauregard MS, Hamel C, St-Arnaud M. Long-term phosphorus fertilization impacts soil fungal and bacterial diversity but not AM fungal community in alfalfa. Microbial Ecology. 2010 Feb 1;59(2):379-89.
Beijerinck MW. Die Bacterien der Papilionaceenkn6llchen. Bmutisc~e Zeitung, 1888 46:725-804.
Bent SJ, Forney LJ: The tragedy of the uncommon: understanding limitations in the analysis of microbial diversity. ISME J 2008, 2:689–695.
Berg G, Smalla K. Plant species and soil type cooperatively shape the structure and function of microbial communities in the rhizosphere. FEMS microbiology ecology. 2009 Apr 1;68(1):1-3.
Berger WH, Parker FL: Diversity of Planktonic Foraminifera in deep-sea sediments. Science. 1970, 168: 1345-1347.
Besemer K, Singer G, Quince C, Bertuzzo E, Sloan W, Battin TJ. Headwaters are critical reservoirs of microbial diversity for fluvial networks. Proceedings of the Royal Society of London B: Biological Sciences. 2013 Nov 22;280(1771):20131760.
Beukema H, Danielsen F, Vincent G, Hardiwinoto S, Van Andel J. Plant and bird diversity in rubber agroforests in the lowlands of Sumatra, Indonesia. Agroforestry Systems. 2007 Jul 1;70(3):217-42.
Blum MG, François O: Which random processes describe the tree of life? A large-scale study of phylogenetic tree imbalance. Syst Biol 2006, 55:685–691.
Bokulich NA, Lewis ZT, Boundy-Mills K, Mills DA. A new perspective on microbial landscapes within food production. Current opinion in biotechnology. 2016 Feb 29;37:182-9.
Bork P, Bowler C, de Vargas C, Gorsky G, Karsenti E, Wincker P. Tara Oceans studies plankton at planetary scale. Science. 2015 May 22;348(6237):873-.
Bowler C, Allen AE, Badger JH, Grimwood J, Jabbari K, Kuo A, Maheswari U, Martens C, Maumus F, Otillar RP, Rayko E. The Phaeodactylum genome reveals the evolutionary history of diatom genomes. Nature. 2008 Nov 13;456(7219):239-44.
Broeckling CD, Broz AK, Bergelson J, Manter DK, Vivanco JM. Root exudates regulate soil fungal community composition and diversity. Applied and environmental microbiology. 2008 Feb 1;74(3):738-44.
Brum JR. 2015. Patterns of ecological drivers of ocean viral communities. Science 348(6237):10.1126/science.1261498.
Brumfield RT, Tello JG, Cheviron ZA, Carling MD, Crochet N, Rosenberg KV: Phylogenetic conservatism and antiquity of a tropical specialization: Army-ant-following in the typical antbirds (Thamnophilidae). Mol Phylogenet Evol 2007, 45:1–13.
Buée M, Reich M, Murat C, Morin E, Nilsson RH, Uroz S, Martin F: 454 Pyrosequencing analyses of forest soils reveal an unexpectedly high fungal diversity. New Phytol 2009, 184:449–456.
Cadotte MW, Davies TJ, Regetz J, Kembel SW, Cleland E, Oakley TH: Phylogenetic diversity metrics for ecological communities: integrating species richness, abundance, and evolutionary history. Ecol Lett 2010, 13:96–105.
Caporaso JG, Bittinger K, Bushman FD, DeSantis TZ, Andersen GL, Knight R. 2010. PyNAST: a flexible tool for aligning sequences to a template alignment. Bioinformatics. 26:266-267.
Caporaso JG, Kuczynski J, Stombaugh J, Bittinger K, Bushman FD, Costello EK, Fierer N, Peña AG, Goodrich JK, Godron JI, Huttley GA, Kelley ST, Knights D, Koenig JE, Ley RE,

Lozupone CA, McDonald D, Muegge BD, Pirrung M, Reeder J, Sevinsky JR, Turnbaugh PJ, Walters WA, Widmann J, Yatsunenko T, Zaneveld J, Knight R: QIIME allows analysis of high- throughput community sequencing data. Nat Methods 2010, 7:335–336.
Carson JK, Gonzalez-Quiñones V, Murphy DV, Hinz C, Shaw JA, Gleeson DB. Low pore connectivity increases bacterial diversity in soil. Applied and environmental microbiology. 2010 Jun 15;76(12):3936-42.
Cermeño P, Falkowski PG, Romero OE, Schaller MF, Vallina SM. Continental erosion and the Cenozoic rise of marine diatoms. Proceedings of the National Academy of Sciences. 2015 Apr 7;112(14):4239-44.
Chan CX, Soares MB, Bonaldo MF, Wisecaver JH, Hackett JD, Anderson DM, Erdner DL, Bhattacharya D. Analysis of Alexandrium tamarense (Dinophyceae) genes reveals the complex evolutionary history of a microbial eukaryote. Journal of phycology. 2012 Oct 1;48(5):1130-42.
Chao A, Chiu C-H, Jost L: Phylogenetic diversity measures based on Hill numbers. Philos Trans R Soc Lond B Biol Sci 2010, 365:3599–3609.
Chaudhary N, Sharma AK, Agarwal P, Gupta A, Sharma VK. 16S Classifier: A Tool for Fast and Accurate Taxonomic Classification of 16S rRNA Hypervariable Regions in Metagenomic Datasets. PloS one. 2015 Feb 3;10(2):e0116106.
Chazdon RL, Careaga S, Webb C, Vargas O: Community and phylogenetic structure of reproductive traits of woody species in wet tropical forests. Ecol Monogr 2003, 73:331–348.
Chevenet F, Brun C, Bañuls AL, Jacq B, Christen R. 2006. TreeDyn: towards dynamic graphics and annotations for analyses of trees. BMC Bioinformatics. 7:439.
Chevenet F, Croce O, Hebrard M, Christen R, Berry V. 2010. ScripTree: scripting phylogenetic graphics. Bioinformatics. 26(8):1125-1126.
Cleveland CC, Reed SC, Keller AB, Nemergut DR, O'Neill SP, Ostertag R, Vitousek PM. Litter quality versus soil microbial community controls over decomposition: a quantitative analysis. Oecologia. 2014 Jan 1;174(1):283-94.
Cleveland CC, Townsend AR, Schmidt SK. Phosphorus limitation of microbial processes in moist tropical forests: evidence from short-term laboratory incubations and field studies. Ecosystems. 2002 Nov 1;5(7):0680-91.
Colless DH: Review of Phylogenetics: The Theory and Practice of Phylogenetic Systematics. Syst Zool 1982, 31:100–104.
Corley RH, Tinker PB. Care and maintenance of oil palms. The oil palm. 2003;4:287-326.
Corley RH. How much palm oil do we need?. Environmental Science & Policy. 2009 Apr 30;12(2):134-9.
Davidson EA, Matson PA, Vitousek PM, Riley R, Dunkin K, Garcia-Mendez G, Maass JM. Processes Regulating Soil Emissions of NO and N^ 2O in a Seasonally Dry Tropical Forest. Ecology. 1993 Jan;74(1):130-9.
De Groot RS, Wilson MA, Boumans RM. A typology for the classification, description and valuation of ecosystem functions, goods and services. Ecological economics. 2002 Jun 30;41(3):393-408.
de Vries FT, Thébault E, Liiri M, Birkhofer K, Tsiafouli MA, Bjørnlund L, Jørgensen HB, Brady MV, Christensen S, de Ruiter PC, d'Hertefeldt T. Soil food web properties explain ecosystem services across European land use systems. Proceedings of the National Academy of Sciences. 2013 Aug 27;110(35):14296-301.

Dismukes GC, Carrieri D, Bennette N, Ananyev GM, Posewitz MC. 2008. Aquatic phototrophs: efficient alternatives to land-based crops for biofuels. Current Opinion in Biotechnology. 19(3): 235-240.
Doll HM, Armitage DW, Daly RA, Emerson JB, Goltsman DS, Yelton AP, Kerekes J, Firestone MK, Potts MD. Utilizing novel diversity estimators to quantify multiple dimensions of microbial biodiversity across domains. BMC microbiology. 2013 Nov 15;13(1):1.
Dominguez-Bello MG, De Jesus-Laboy KM, Shen N, Cox LM, Amir A, Gonzalez A, Bokulich NA, Song SJ, Hoashi M, Rivera-Vinas JI, Mendez K. Partial restoration of the microbiota of cesarean-born infants via vaginal microbial transfer. Nature medicine. 2016 Feb 1.
Dornelas M, Gotelli NJ, McGill B, Shimadzu H, Moyes F, Sievers C, Magurran AE. Assemblage time series reveal biodiversity change but not systematic loss. Science. 2014 Apr 18;344(6181):296-9.
Drenovsky RE, Vo D, Graham KJ, Scow KM. Soil water content and organic carbon availability are major determinants of soil microbial community composition. Microbial Ecology. 2004 Nov 1;48(3):424-30.
Edgar RC: MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Res 2004, 32:1792–1797.
Edgar RC: Search and clustering orders of magnitude faster than BLAST. Bioinf 2010, 26:2460–2461.
Eisen JA. Horizontal gene transfer among microbial genomes: new insights from complete genome analysis. Current opinion in genetics & development. 2000 Dec 1;10(6):606-11.
Emerson JB, Andrade K, Thomas BC, Norman A, Allen EE, Heidelberg KB, Banfield JF: Virus-host and CRISPR dynamics in archaea-dominated Hypersaline Lake Tyrrell, Victoria, Australia. Archaea 2013A, 2013:370871.
Emerson JB, Thomas BC, Andrade K, Allen EE, Heidelberg KB, Banfield JF: Dynamic viral populations in hypersaline systems as revealed by metagenomic assembly. Appl Environ Microbiol 2012, 78:6309–20.
Emerson JB, Thomas BC, Andrade K, Heidelberg KB, Banfield JF. New approaches indicate constant viral diversity despite shifts in assemblage structure in an Australian hypersaline lake. Applied and environmental microbiology. 2013B Nov 1;79(21):6755-64.
Faith DP: Conservation evaluation and phylogenetic diversity. Biol Conserv 1992, 61:1–10.
Falciatore A, Bowler C. 2002. Revealing the molecular secrets of marine diatoms. Annual Review of Plant Biology. 53:109-130.
Fearnside PM. Global warming and tropical land-use change: greenhouse gas emissions from biomass burning, decomposition and soils in forest conversion, shifting cultivation and secondary vegetation. Climatic change. 2000 Jul 1;46(1-2):115-58.
Ferrer M, Beloqui A, Timmis KM, Golyshin KN: Metagenomics for mining new genetic resources of microbial communities. J Mol Microbiol Biotechnol 2009, 16:109–123.
Field CB, Behrenfeld MJ, Randerson JT, Falkowski P. 1998. Primary production of the biosphere: integrating terrestrial and oceanic components. Science. 281:237–240.
Fierer N, McCain CM, Meir P, Zimmermann M, Rapp JM, Silman MR, Knight R: Microbes do not follow the elevational diversity patterns of plants and animals. Ecology 2011, 92:797–804.
Finkel ZV, Katz ME, Wright JD, Schofield OM, Falkowski PG. Climatically driven macroevolutionary patterns in the size of marine diatoms over the Cenozoic. Proceedings

of the National Academy of Sciences of the United States of America. 2005 Jun 21;102(25):8927-32.
Fisher RA, Corbet AS, Williams CB: The Relation Between the Number of Species and the Number of Individuals in a Random Sample of an Animal Population. J Anim Ecol 1943, 12:42–58.
Frey SD, Knorr M, Parrent JL, Simpson RT. Chronic nitrogen enrichment affects the structure and function of the soil microbial community in temperate hardwood and pine forests. Forest Ecology and Management. 2004 Jul 12;196(1):159-71.
Fritsch FE. 1935. Structure and reproduction of the algae. Vol. 1. Cambridge University Press. Cambridge.
Fulthorpe RR, Roesch LFW, Riva A, Triplett EW: Distantly sampled soils carry few species in common. ISME J 2008, 2:901–910.
Galvão TC, Mohn WW, de Lorenzo V: Exploring the microbial biodegradation and biotransformation gene pool. Trends Biotechnol 2005, 23:497–506.
Gibbs HK, Ruesch AS, Achard F, Clayton MK, Holmgren P, Ramankutty N, Foley JA. Tropical forests were the primary sources of new agricultural land in the 1980s and 1990s. Proceedings of the National Academy of Sciences. 2010 Sep 21;107(38):16732-7.
Goltsman D: Community Genomic, Proteomic, and Transcriptomic Analyses of Acid Mine Drainage Biofilm Communities, PhD thesis. Berkeley, California, USA: University of California Berkeley, Environmental Science, Policy and Management Department; 2013.
Green J, Bohannan BJ. Spatial scaling of microbial biodiversity. Trends in ecology & evolution. 2006 Sep 30;21(9):501-7.
Griffin DM. Soil moisture and the ecology of soil fungi. Biological Reviews. 1963 May 1;38(2):141-66.
Griffiths E, Gupta RS. Identification of signature proteins that are distinctive of the" Deinococcus-Thermus" phylum. International microbiology: official journal of the Spanish Society for Microbiology. 2007;10(3):201-8.
Guillaume T, Damris M, Kuzyakov Y. 2015. Losses of soil carbon by converting tropical forest to plantations: erosion and decomposition estimated by $\delta^{13}C$. Global Change Biology. 21(9):3548-3560.
Guillou L, Bachar D, Audic S, Bass D, Berney C, Bittner L, et al. 2013. The Protist Ribosomal Reference database (PR2): a catalog of unicellular eukaryote Small Sub-Unit rRNA sequences with curated taxonomy. Nucleic Acids Res. 41(D1):D597-D604.
Haeckel EH. Generelle Morphologie der Organismen allgemeine Grundzuge der organischen Formen-Wissenschaft, mechanisch begrundet durch die von Charles Darwin reformirte Descendenz-Theorie von Ernst Haeckel: Allgemeine Entwickelungsgeschichte der Organismen kritische Grundzuge der mechanischen Wissenschaft von den entstehenden Formen der Organismen, begrundet durch die Descendenz-Theorie. Verlag von Georg Reimer; 1866.
Haegeman B, Hamelin J, Moriaty J, Nael P, Dushoff J, Weitz JS: Robust estimation of microbial diversity in theory and in practice. ISME J 2013. doi:10.1038/ismej.2013.10.
Hagen JB. An entangled bank: the origins of ecosystem ecology. Rutgers University Press; 1992.
Hansen MC, Potapov PV, Moore R, Hancher M, Turubanova SA, Tyukavina A, Thau D, Stehman SV, Goetz SJ, Loveland TR, Kommareddy A. High-resolution global maps of 21st-century forest cover change. science. 2013 Nov 15;342(6160):850-3.

Harmon Luke J, Jason T Weir, Chad D Brock, Richard E Glor, and Wendell Challenger. 2008. GEIGER: investigating evolutionary radiations. Bioinformatics 24:129-131.

Hartmann A, Schmid M, Van Tuinen D, Berg G. Plant-driven selection of microbes. Plant and Soil. 2009 Aug 1;321(1-2):235-57.

Hasle GR. 1974. The 'mucilage pore' of pennate diatoms. Nova Hedwigia, Beiheft. 45: 167-194.

Hewitt CN, MacKenzie AR, Di Carlo P, Di Marco CF, Dorsey JR, Evans M, Fowler D, Gallagher MW, Hopkins JR, Jones CE, Langford B. Nitrogen management is essential to prevent tropical oil palm plantations from causing ground-level ozone pollution. Proceedings of the National Academy of Sciences. 2009 Nov 3;106(44):18447-51.

Hill MO: Diversity and Evenness: A Unifying Notation and Its Consequences. Ecology 1973, 54:427–432.

Hill TCJ, Walsh KA, Harris JA, Moffett BF: Using ecological diversity measures with bacterial communities. FEMS Microbiol Ecol 2003, 43:1–11.

Hooper DU, Vitousek PM: The effects of plant composition and diversity on ecosystem processes. Science 1997, 277:1302–1305.

Horner-Devine MC, Lage M, Hughes JB, Bohannan BJM: A taxa-area relationship for bacteria. Nature 2004, 432:750–753.

Houghton RA. How well do we know the flux of CO2 from land-use change?. Tellus B. 2010 Nov 1;62(5):337-51.

Hughes JB, Hellmann JJ, Ricketts TH, Bohannan BJ. Counting the uncountable: statistical approaches to estimating microbial diversity. Applied and environmental microbiology. 2001 Oct 1;67(10):4399-406.

Hyatt D, Chen G-L, LoCascio PF, Land ML, Larimer FW, Hauser LJ: Prodigal: prokaryotic gene recognition and translation initiation site identification. BMC Bioinf 2010, 11:119.

Illumina. "HiSeq 3000/HiSeq 4000 Sequencing systems". http://www.illumina.com/systems/hiseq-3000-4000.html. Accessed June 25, 2016.

Isbell F, Calcagno V, Hector A, Connolly J, Harpole WS, Reich PB, Scherer-Lorenzen M, Schmid B, Tilman D, van Ruijven J, Weigelt A. High plant diversity is needed to maintain ecosystem services. Nature. 2011 Sep 8;477(7363):199-202.

Jaccard P. Distribution de la flore alpine dans le basin de dranses et dans quelques regions voisines. Bull Société Vaudoise Sci Natur 1901, 37:241272.

Jawjit W, Kroeze C, Rattanapan S. Greenhouse gas emissions from rubber industry in Thailand. Journal of cleaner production. 2010 Mar 31;18(5):403-11.

Jetz W, Thomas GH, Joy JB, Hartmann K, Mooers AO. 2012. The global diversity of birds in space and time. Nature. 491:444-448.

Jost L: Entropy and diversity. Oikos 2006, 113:363–375.

Jousset A, Eisenhauer N, Materne E, Scheu S. Evolutionary history predicts the stability of cooperation in microbial communities. Nature communications. 2013 Oct 11;4.

Kembel SW, Cowan PD, Helmus MR, Cornwell WK, Morlon H, Ackerly DD, Blomberg SP, Webb CO. Picante: R tools for integrating phylogenies and ecology. Bioinformatics. 2010 Jun 1;26(11):1463-4.

Kembel SW, Jones E, Kline J, Northcutt D, Stenson J, Womack AM, Bohannan BJ, Brown GZ, Green JL. Architectural design influences the diversity and structure of the built environment microbiome. The ISME journal. 2012 Aug 1;6(8):1469-79.

Kerekes J: Species Diversity, Ecology and Laccase Gene Diversity of Saprotrophic Fungi across Different Plant Community Types, PhD thesis. Berkeley, California, USA: University of California, Berkeley, Department of Plant and Microbial Biology; 2011.
Koh LP, Wilcove DS. Is oil palm agriculture really destroying tropical biodiversity?. Conservation letters. 2008 Jun 1;1(2):60-4.
Kooistra WHCF, Forlani G, Stefano MD. 2009. Adaptation of araphid pennate diatoms to a planktonic existence. Marine Ecology. 30:1-15.
Kooistra WHCF, Gersonde R, Medlin LK, Mann DG. 2007. The Origin and Evolution of Diatoms: Their Adaptation to a Planktonic Existence. In Evolution of Primary Producers in the Sea. Burlington, MA. Pp 210-241.
Kruys N, Jonsson BG. Fine woody debris is important for species richness on logs in managed boreal spruce forests of northern Sweden. Canadian Journal of Forest Research. 1999 Sep 1;29(8):1295-9.
Landis FC, Gargas A, Givnish TJ. Relationships among arbuscular mycorrhizal fungi, vascular plants and environmental conditions in oak savannas. New Phytologist. 2004 Dec 1;164(3):493-504.
Langille MGI, Zaneveld J, Caporaso JG, McDonald D, Knights D, Reyes JA, Celemente JC, Burkepile DE, Vega Thurber RL, Knight R, Beiko RG, Huttenhower C: Predictive functional profiling of microbial communities using 16S rRNA marker gene sequences. Nature Biotechnology 2013, 31:814–821.
Lauber CL, Hamady M, Knight R, Fierer N. Pyrosequencing-based assessment of soil pH as a predictor of soil bacterial community structure at the continental scale. Applied and environmental microbiology. 2009 Aug 1;75(15):5111-20.
Laurance WF, Useche DC, Rendeiro J, Kalka M, Bradshaw CJ, Sloan SP, Laurance SG, Campbell M, Abernethy K, Alvarez P, Arroyo-Rodriguez V. Averting biodiversity collapse in tropical forest protected areas. Nature. 2012 Sep 13;489(7415):290-4.
Lazarus D, Barron J, Renaudie J, Diver P, Türke A. Cenozoic planktonic marine diatom diversity and correlation to climate change. PloS one. 2014 Jan 22;9(1):e84857.
Leblanc K, Arístegui J, Kopczynska E, Marshall H, Peloquin J, Piontkovski S, Poulton AJ, Quéguiner B, Schiebel R, Shipe R, Stefels J. A global diatom database–abundance, biovolume and biomass in the world ocean.
Lee-Cruz L, Edwards DP, Tripathi BM, Adams JM. Impact of logging and forest conversion to oil palm plantations on soil bacterial communities in Borneo. Applied and environmental microbiology. 2013 Dec 1;79(23):7290-7.
Leinster T, Cobbold CA. Measuring diversity: the importance of species similarity. Ecology. 2012 Mar;93(3):477-89.
Levitan O, Dinamarca J, Hochman G, Falkowski PG. 2014. Diatoms: a fossil fuel of the future. Trends in Biotechnology. 32(3): 117-124.
Lima-Mendez G, Faust K, Henry N, Decelle J, Colin S, Carcillo F, Chaffron S, Ignacio-Espinosa JC, Roux S, Vincent F, Bittner L. Determinants of community structure in the global plankton interactome. Science. 2015 May 22;348(6237):1262073.
Liu L, Gundersen P, Zhang T, Mo J. Effects of phosphorus addition on soil microbial biomass and community composition in three forest types in tropical China. Soil Biology and Biochemistry. 2012 Jan 31;44(1):31-8.
Lozupone C, Knight R: UniFrac: a New Phylogenetic Method for Comparing Microbial Communities. Micrbiol 2005, 71:8228–8235.

Luis P, Kellner H, Zimdars B, Langer U, Martin F, Buscot F. Patchiness and spatial distribution of laccase genes of ectomycorrhizal, saprotrophic, and unknown basidiomycetes in the upper horizons of a mixed forest cambisol. Microbial Ecology. 2005 Nov 1;50(4):570-9.
Luo F, Yang Y, Zhong J, Gao H, Khan L, Thompson DK, Zhou J: Constructing gene co-expression networks and predicting functions of unknown genes by random matrix theory. BMC Bioinf 2007, 8:299.
Luskin MS, Potts MD. Microclimate and habitat heterogeneity through the oil palm lifecycle. Basic and Applied Ecology. 2011 Sep 30;12(6):540-51.
Magurran AE, Henderson PA: Explaining the excess of rare species in natural species abundance distributions. Nature 2003, 422:714–716.
Mahé F, Rognes T, Quince C, de Vargas C, Dunthorn M. 2014. Swarm: robust and fast clustering method for amplicon-based studies. Peer J. 2:e593 https://doi.org/10.7717/peerj.593.
Mardis ER. A decade/'s perspective on DNA sequencing technology. Nature. 2011 Feb 10;470(7333):198-203.
Margulies M, Egholm M, Altman WE, Attiya S, Bader JS, Bemben LA, Berka J, Braverman MS, Chen Y-J, Chen Z, Dewell SB, Du L, Fierro JM, Gomes XV, Godwin BC, He W, Helgesen S, Ho CH, Irzyk GP, Jando SC, Alenquer MLI, Jarvie TP, Jirage KB, Kim J-B, Knight JR, Lanza JR, Leamon JH, Lefkowitz SM, Lei M, Li J, et al: Genome sequencing in microfabricated high-density picolitre reactors. Nature 2005, 437:376–380.
Martiny AC, Treseder K, Pusch G: Phylogenetic conservatism of functional traits in microorganisms. ISME J 2013, 7:830–838.
May MR, Moore BR. 2016. How well can we detect lineage-specific diversification-rate shifts? A simulation study of sequential AIC methods. Syst Bio. 10.1093/sysbio/syw026.
McGuire KL, D'Angelo H, Brearley FQ, Gedallovich SM, Babar N, Yang N, Gillikin CM, Gradoville R, Bateman C, Turner BL, Mansor P. Responses of soil fungi to logging and oil palm agriculture in Southeast Asian tropical forests. Microbial ecology. 2015 May 1;69(4):733-47.
Medlin LK, Williams DM, Sims PA. 1993. The evolution of the diatoms (Bacillariophyta). I. Origin of the group and assessment of the monophyly of its major divisions. European J Phycology. 28(4):261-275.
Melling L, Hatano R, Goh KJ. Methane fluxes from three ecosystems in tropical peatland of Sarawak, Malaysia. Soil Biology and Biochemistry. 2005a Aug 31;37(8):1445-53.
Melling L, Hatano R, Goh KJ. Soil CO2 flux from three ecosystems in tropical peatland of Sarawak, Malaysia. Tellus B. 2005b Feb 1;57(1):1-1.
Melo FP, Arroyo-Rodríguez V, Fahrig L, Martínez-Ramos M, Tabarelli M. On the hope for biodiversity-friendly tropical landscapes. Trends in ecology & evolution. 2013 Aug 31;28(8):462-8.
Mendes LW, de Lima Brossi MJ, Kuramae EE, Tsai SM. Land-use system shapes soil bacterial communities in Southeastern Amazon region. Applied soil ecology. 2015 Nov 30;95:151-60.
Miller CS, Baker BJ, Thomas BC, Singer SW, Banfield JF: EMIRGE: reconstruction of full-length ribosomal genes from microbial community short read sequencing data. Genome Biol 2011, 12:R44.

Miller CS, Handley KM, Wrighton KC, Frischkorn KR, Thomas BC, Banfield JF: Short-Read Assembly of Full-Length 16S Amplicons Reveals Bacterial Diversity in Subsurface Sediments. PLoS ONE 2013, 8(2):e56018. doi: 10.1371/journal.pone.0056018.

Mirarab S, Nguyen N, Guo S, Wang LS, Kim J, Warnow T. PASTA: ultra-large multiple sequence alignment for nucleotide and amino-acid sequences. Journal of Computational Biology. 2015 May 1;22(5):377-86.

Mishler BD, Knerr N, González-Orozco CE, Thornhill AH, Laffan SW, Miller JT. Phylogenetic measures of biodiversity and neo-and paleo-endemism in Australian Acacia. Nature Communications. 2014 Jul 18;5.

Mishler BD: Species are not uniquely real biological entities. In Contemporary Debates in Philosophy of Biology. Edited by Ayala FJ, Arp R. Oxford: Wiley-Blackwell; 2010:110–122.

Monroy F, van der Putten WH, Yergeau E, Mortimer SR, Duyts H, Bezemer TM. Community patterns of soil bacteria and nematodes in relation to geographic distance. Soil Biology and Biochemistry. 2012 Feb 29;45:1-7.

Mooers AØ, Heard SB: Inferring evolutionary process from phylogenetic tree shape. Q Rev Biol 1997, 72:31–54.

Morlon H, Parsons TL, Plotkin JB. Reconciling molecular phylogenies with the fossil record. Proceedings of the National Academy of Sciences. 2011 Sep 27;108(39):16327-32.

Morlon H. Phylogenetic approaches for studying diversification. Ecology letters. 2014 Apr 1;17(4):508-25.

Myers N, Mittermeier RA, Mittermeier CG, Da Fonseca GA, Kent J. Biodiversity hotspots for conservation priorities. Nature. 2000 Feb 24;403(6772):853-8.

Nawrocki EP, Kolbe DL, Eddy SR: Infernal 1.0: inference of RNA alignments. Bioinf 2009, 25:1335–1337.

Near TJ, Dornburg A, Eytan RI, Keck BP, Smith WL, et al. 2013. Phylogeny and tempo of diversification in the superradiation of spiny-rayed fishes. Proc Nat Acad Sci. 110(31)L12738:12743.

Nelson DM, Treguer P, Brzezinski MA, Leynaert A, Queguiner B. 1995. Production and dissolution of biogenic silica in the ocean: revised global estimates, comparison with regional data and relationship to biogenic sedimentation. Glob. Biogeochem. Cycles. 9: 359–372.

Nelson E, Mendoza G, Regetz J, Polasky S, Tallis H, Cameron D, Chan K, Daily GC, Goldstein J, Kareiva PM, Lonsdorf E. Modeling multiple ecosystem services, biodiversity conservation, commodity production, and tradeoffs at landscape scales. Frontiers in Ecology and the Environment. 2009 Feb 1;7(1):4-11.

Newbold T, Hudson LN, Phillips HRP, Hill SLL, Contu S, et al. 2014. A global model of the response of tropical and sub-tropical forest biodiversity to anthropogenic pressures. Proc Royal Society B. 281(1792):20141371.

Nielsen UN, Ayres E, Wall DH, Bardgett RD. Soil biodiversity and carbon cycling: a review and synthesis of studies examining diversity–function relationships. European Journal of Soil Science. 2011 Feb 1;62(1):105-16.

O'Brien HE, Parrent JL, Jackson JA, Moncalvo J-M, Vilgalys R: Fungal community analysis by large-scale sequencing of environmental samples. Appl Environ Microbiol 2005, 71:5544–5550.

Oksanen J, Blanchet FG, Kindt R, Legendre P, Minchin RR, O'Hara RB, Simpson GL, Solymos P, Stevens MHH, Wagner H. vegan: Community Ecology Package. 2016. R package version 2.3-4.
Paradis E, Claude J, Strimmer K. APE: analyses of phylogenetics and evolution in R language. Bioinformatics 2004. 20: 289-290.
Parsons WF, Mitre ME, Keller M, Reiners WA. Nitrate limitation of N2O production and denitrification from tropical pasture and rain forest soils. Biogeochemistry. 1993 Jan 1;22(3):179-93.
Paula FS, Rodrigues JL, Zhou J, Wu L, Mueller RC, Mirza BS, Bohannan BJ, Nüsslein K, Deng Y, Tiedje JM, Pellizari VH. Land use change alters functional gene diversity, composition and abundance in Amazon forest soil microbial communities. Molecular ecology. 2014 Jun 1;23(12):2988-99.
Peay KG, Baraloto C, Fine PV. Strong coupling of plant and fungal community structure across western Amazonian rainforests. The ISME journal. 2013 Sep 1;7(9):1852-61.
Pedrós-Alió C. Marine microbial diversity: can it be determined?. Trends in microbiology. 2006 Jun 30;14(6):257-63.
Pershina EV, Andronov EE, Pinaev AG, Provorov NA. Recent advances and perspectives in metagenomic studies of soil microbial communities. InManagement of Microbial Resources in the Environment 2013 (pp. 141-166). Springer Netherlands.
Pesant S, Not F, Picheral M, Kandels-Lewis S, Le Bescot N, Gorsky G, Iudicone D, Karsenti E, Speich S, Troublé R, Dimier C. Open science resources for the discovery and analysis of Tara Oceans data. Scientific data. 2015 May 26;2.
Placella SA, Brodie EL, Firestone MK: Rainfall-induced carbon dioxide pulses result from sequential resuscitation of phylogenetically clustered microbial groups. Proc Natl Acad Sci U S A 2012, 109:10931–10936.
Postel SL, Thompson BH. Watershed protection: Capturing the benefits of nature's water supply services. InNatural Resources Forum 2005 May 1 (Vol. 29, No. 2, pp. 98-108). Blackwell Publishing, Ltd..
Price MN, Dehal PS, Arkin AP: FastTree 2 – Approximately Maximum-Likelihood Trees for Large Alignments. PLoS ONE 2010, 5:e9490. Doi: 10.1371/journal.pone.0009490.
Price MN, Dehal PS, Arkin AP: FastTree: computing large minimum evolution trees with profiles instead of a distance matrix. Mol Biol Evol 2009, 26:1641–1650.
Quevillon E, Silventoinen V, Pillai S, Harte N, Mulder N, Apweiler R, Lopez R: InterProScan: protein domains identifier. Nucleic Acids Res 2005, 33:W116–W120.
Quince C, Curtis TP, Sloan WT. 2008. The rational exploration of microbial diversity. 2008. The ISME Journal. 2,:997–1006.
R Core Team. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. 2014.
Rabosky DL, Sorhannus U. Diversity dynamics of marine planktonic diatoms across the Cenozoic. Nature. 2009 Jan 8;457(7226):183-6.
Ramage BS, Sheil D, Salim HM, Fletcher C, MUSTAFA NZ, Luruthusamay JC, Harrison RD, Butod E, Dzulkiply AD, Kassim AR, Potts MD. Pseudoreplication in tropical forests and the resulting effects on biodiversity conservation. Conservation Biology. 2013 Apr 1;27(2):364-72.
Rashid M, Stingl U. Contemporary molecular tools in microbial ecology and their application to advancing biotechnology. Biotechnology advances. 2015 Dec 31;33(8):1755-73.

Reeve JR, Schadt CW, Carpenter-Boggs L, Kang S, Zhou J, Reganold JP. Effects of soil type and farm management on soil ecological functional genes and microbial activities. The ISME journal. 2010 Sep 1;4(9):1099-107.
Ricklefs RE. Environmental heterogeneity and plant species diversity: a hypothesis. The American Naturalist. 1977 Mar 1;111(978):376-81.
Rizzo L, Manaia C, Merlin C, Schwartz T, Dagot C, Ploy MC, Michael I, Fatta-Kassinos D. Urban wastewater treatment plants as hotspots for antibiotic resistant bacteria and genes spread into the environment: a review. Science of the total environment. 2013 Mar 1;447:345-60.
Roberts A, Pimentel H, Trapnell C, Pachter L: Identification of novel transcripts in annotated genomes using RNA-Seq. Bioinf 2011, 27:2325–2329.
Robertson GP, Grace PR. Greenhouse gas fluxes in tropical and temperate agriculture: the need for a full-cost accounting of global warming potentials. InTropical Agriculture in Transition—Opportunities for Mitigating Greenhouse Gas Emissions? 2004 (pp. 51-63). Springer Netherlands.
Rodrigues JL, Pellizari VH, Mueller R, Baek K, Jesus ED, Paula FS, Mirza B, Hamaoui GS, Tsai SM, Feigl B, Tiedje JM. Conversion of the Amazon rainforest to agriculture results in biotic homogenization of soil bacterial communities. Proceedings of the National Academy of Sciences. 2013 Jan 15;110(3):988-93.
Roesch LFW, Fulthorpe RR, Riva A, Casella G, Hadwin AKM, Kent AD, Daroub SH, Camargo FAO, Farmerie WG, Triplett EW: Pyrosequencing enumerates and contrasts soil microbial diversity. ISME J 2007, 1:283–290.
Rosselló-Mora R, Amann R: The species concept for prokaryotes. FEMS Microbiol Rev 2001, 25:39–67.
Sanderson, M. J. (2002) Estimating absolute rates of molecular evolution and divergence times: a penalized likelihood approach. Molecular Biology and Evolution, 19, 101–109.
Sanmartín I, Meseguer AS. Extinction in Phylogenetics and Biogeography: From Timetrees to Patterns of Biotic Assemblage. Frontiers in genetics. 2016;7.
Sarthou G, Timmermans KR, Blain S, Treguer P. 2005. Growth physiology and fate of diatoms in the ocean: a review. J. Sea Res. 53:25-42.
Schneider D, Engelhaupt M, Allen K, Kurniawan S, Krashevska V, Heinemann M, Nacke H, Wijayanti M, Meryandini A, Corre MD, Scheu S. Impact of Lowland Rainforest Transformation on Diversity and Composition of Soil Prokaryotic Communities in Sumatra (Indonesia). Frontiers in microbiology. 2015;6.
Sessitsch A, Weilharter A, Gerzabek MH, Kirchmann H, Kandeler E. Microbial population structures in soil particle size fractions of a long-term fertilizer field experiment. Applied and Environmental Microbiology. 2001 Sep 1;67(9):4215-24.
Setälä H, McLean MA. Decomposition rate of organic substrates in relation to the species diversity of soil saprophytic fungi. Oecologia. 2004 Mar 1;139(1):98-107.
Shannon CE: A Mathematical Theory of Communication. Bell System Technical Journal 1948, 27:379–423.
Silver WL, Thompson AW, Reich A, Ewel JJ, Firestone MK. Nitrogen cycling in tropical plantation forests: potential controls on nitrogen retention. Ecological Applications. 2005 Oct;15(5):1604-14.
Silvertown J: Plant coexistence and the niche. Trends Ecol Evol 2004, 19:605–611.
Simonsen R. 1979. The diatom system: ideas on phylogeny. Bacillaria. 2:9–71.

Simpson EH: Meaasurement of diversity. Nature. 1949, 163: 688-10.1038/163688a0.
Simpson JT, Wong K, Jackman SD, Schein JE, Jones SJM, Birol I: ABySS: A parallel assembler for short read sequence data. Genome Res 2009, 19:1117–1123.
Sims PA, Mann DG, Medlin LK. 2006. Evolution of the diatoms: insights from fossil, biological and molecular data. Phycologia. 45(4):361-402.
Singh BK, Macdonald CA: Drug discovery from uncultivable microorganisms. Drug Discov Today 2010, 15:792–799.
Sodhi NS, Koh LP, Brook BW, Ng PK. Southeast Asian biodiversity: an impending disaster. Trends in Ecology & Evolution. 2004 Dec 31;19(12):654-60.
Sogin ML, Morrison HG, Huber JA, Welch DM, Huse SM, Neal PR, Arrieta JM, Herndl GJ: Microbial diversity in the deep sea and the underexplored "rare biodsphere. Proc Natl Acad Sci U S A 2006, 103:12115–12120.
Staley JT: The bacterial species dilemma and the genomic-phylogenetic species concept. Philos Trans R Soc Lond B Biol Sci 2006, 361:1899–1909.
Stephens B, Adams RI, Bhangar S, Bibby K, Waring MS. From commensalism to mutualism: integrating the microbial ecology, building science, and indoor air communities to advance research on the indoor microbiome. Indoor air. 2015 Feb 1;25(1):1-3.
Stibig HJ, Achard F, Carboni S, Raši R, Miettinen J. Change in tropical forest cover of Southeast Asia from 1990 to 2010. Biogeosciences. 2014 Jan 22;11(2):247-58.
Sunagawa S, Coelho LP, Chaffron S, Kultima JR, Labadie K, Salazar G, Djahanschiri B, Zeller G, Mende DR, Alberti A, Cornejo-Castillo FM. Structure and function of the global ocean microbiome. Science. 2015 May 22;348(6237):1261359.
Sunagawa S, Woodley CM, Medina M: Threatened Corals Provide Underexplored Microbial Habitats. PLoS ONE 2010, 5:e9554. Doi: 10.1371/journal.pone.0009554.
Taylor JW, Jacobson DJ, Fisher MC. The evolution of asexual fungi: reproduction, speciation and classification. Annual review of phytopathology. 1999 Sep;37(1):197-246.
Taylor JW, Jacobson DJ, Kroken S, Kasuga T, Geiser DM, Hibbett DS, Fisher MC: Phylogenetic species recognition and species concepts in fungi. Fung Genet Biol 2000, 31:21–32.
Tedersoo L, Jairus T, Horton BM, Abarenkov K, Suvi T, Saar I, Kõljalg U: Strong host preference of ectomycorrhizal fungi in a Tasmanian wet sclerophyll forest as revealed by DNA barcoding and taxon-specific primers. New Phytol 2008, 180:479–490.
Theriot EC, Ashworth M, Ruck E, Nakov T, Jansen RK. 2010. A preliminary multigene phylogeny phylogeny of the diatoms (Bacillariophyta): challenges for future research. Plant Ecology and Evolution. 143(3):278-296.
Theriot EC, Cannone JJ, Gutell RR, Alverson AJ. The limits of nuclear-encoded SSU rDNA for resolving the diatom phylogeny. European Journal of Phycology. 2009 Aug 1;44(3):277-90.
Thomsen PF, Willerslev E. Environmental DNA–an emerging tool in conservation for monitoring past and present biodiversity. Biological Conservation. 2015 Mar 31;183:4-18.
Thornhill AH, Mishler BD, Knerr NJ, González-Orozco CE, Costion CM, Crayn DM, Laffan SW, Miller JT. Continental-scale spatial phylogenetics of Australian angiosperms provides insights into ecology, evolution and conservation. Journal of Biogeography. 2016 May 1.
Tiedje JM, Asuming-Brempong S, Nüsslein K, Marsh TL, Flynn SJ: Opening the black box of soil microbial diversity. Appl Soil Ecol 1999, 13:109–122.

Tilman D, Lehman CL, Thomson KT: Plant diversity and ecosystem productivity: Theoretical considerations. Proc Natl Acad Sci U S A 1997, 94:1857–1861.
Torsvik V, Øvreås L. Microbial diversity and function in soil: from genes to ecosystems. Current opinion in microbiology. 2002 Jun 1;5(3):240-5.
Treseder KK, Allen MF. Direct nitrogen and phosphorus limitation of arbuscular mycorrhizal fungi: a model and field test. New Phytologist. 2002 Sep 1;155(3):507-15.
Tripathi BM, Kim M, Lai-Hoe A, Shukor NA, Rahim RA, Go R, Adams JM. pH dominates variation in tropical soil archaeal diversity and community structure. FEMS microbiology ecology. 2013 Nov 1;86(2):303-11.
Tripathi BM, Kim M, Singh D, Lee-Cruz L, Lai-Hoe A, Ainuddin AN, Go R, Rahim RA, Husni MH, Chun J, Adams JM. Tropical soil bacterial communities in Malaysia: pH dominates in the equatorial tropics too. Microbial ecology. 2012 Aug 1;64(2):474-84.
Tripathi BM, Lee-Cruz L, Kim M, Singh D, Go R, Shukor NA, Husni MH, Chun J, Adams JM. Spatial scaling effects on soil bacterial communities in Malaysian tropical forests. Microbial ecology. 2014 Aug 1;68(2):247-58.
Tripathi BM, Song W, Slik JW, Sukri RS, Jaafar S, Dong K, Adams JM. Distinctive tropical forest variants have unique soil microbial communities, but not always low microbial diversity. Frontiers in microbiology. 2016;7.
Valls M, De Lorenzo V. Exploiting the genetic and biochemical capacities of bacteria for the remediation of heavy metal pollution. FEMS Microbiology Reviews. 2002 Nov 1;26(4):327-38.
Van Der Heijden MG, Bardgett RD, Van Straalen NM. The unseen majority: soil microbes as drivers of plant diversity and productivity in terrestrial ecosystems. Ecology letters. 2008 Mar 1;11(3):296-310.
Wallace JR (2012). Imap: Interactive Mapping. R package version 1.32.
Ward JH: Hierarchical Grouping to Optimize an Objective Function. J Am Stat Assoc 1963, 58:236–244.
Watling R, Harper DB. Chloromethane production by wood-rotting fungi and an estimate of the global flux to the atmosphere. Mycological Research. 1998 Jul 1;102(07):769-87.
Werner C, Kiese R, Butterbach-Bahl K. Soil-atmosphere exchange of N2O, CH4, and CO2 and controlling environmental factors for tropical rain forest sites in western Kenya. Journal of Geophysical Research: Atmospheres. 2007 Feb 16;112(D3).
Wieland G, Neumann R, Backhaus H. Variation of microbial communities in soil, rhizosphere, and rhizoplane in response to crop species, soil type, and crop development. Applied and Environmental Microbiology. 2001 Dec 1;67(12):5849-54.
Willner D, Hugenholtz P. From deep sequencing to viral tagging: recent advances in viral metagenomics. Bioessays. 2013 May 1;35(5):436-42.
Winogradsky S. Ber Schwefelbakterien. Botanische Zeitgung. 1887;45.
Worm B, Barbier EB, Beaumont N, Duffy JE, Folke C, Halpern BS, Jackson JB, Lotze HK, Micheli F, Palumbi SR, Sala E. Impacts of biodiversity loss on ocean ecosystem services. science. 2006 Nov 3;314(5800):787-90.
Xu J. Invited review: microbial ecology in the age of genomics and metagenomics: concepts, tools, and recent advances. Molecular ecology. 2006 Jun 1;15(7):1713-31.
Yang Y, Gao Y, Wang S, Xu D, Yu H, Wu L, Lin Q, Hu Y, Li X, He Z, Deng Y. The microbial gene diversity along an elevation gradient of the Tibetan grassland. The ISME journal. 2014 Feb 1;8(2):430-40.

Yelton AP, Williams KH, Fournelle J, Wrighton KC, Handley KM, Banfield JF: Vanadate and acetate biostimulation of contaminated sediments decreases diversity, selects for specific taxa, and decreases aqueous v(5+) concentration. Environ Sci Technol 2013, 47:6500–6509.

Zerbino DR, Birney E: Velvet: Algorithms for de novo short read assembly using de Bruijn graphs. Genome Res 2008, 18:821–829.

Zhang Y, Cong J, Lu H, Yang C, Yang Y, Zhou J, Li D. 2014. An integrated study to analyze soil microbial community structure and metabolic potential in two forest types. PLoS One. 9(4):e93773.

Zhang Y, Lu Z, Liu S, Yang Y, He Z, Ren Z, Zhou J, Li D. Geochip-based analysis of microbial communities in alpine meadow soils in the Qinghai-Tibetan plateau. BMC microbiology. 2013 Mar 29;13(1):1.

Zhou J, Kang S, Schadt CW, Garten CT. Spatial scaling of functional gene diversity across various microbial taxa. Proceedings of the National Academy of Sciences. 2008 Jun 3;105(22):7768-73.

Ziegler AD, Fox JM, Xu J. The rubber juggernaut. Science. 2009 May 22;324(5930):1024-5.