Method development for the analysis of soil bacterial communities

by

Andrea Kate Bartram

A thesis
presented to the University of Waterloo
in fulfillment of the
thesis requirement for the degree of
Doctor of Philosophy
in
Biology

Waterloo, Ontario, Canada, 2014

AUTHOR'S DECLARATION

This thesis consists of material all of which I authored or co-authored: see Statement of Contributions included in the thesis. This is a true copy of the thesis, including any required final revisions, as accepted by my examiners.

I understand that my thesis may be made electronically available to the public.

SATEMENT OF CONTRIBUTIONS

Figure 2.2 and 2.7 were produced by Michael Lynch and NMF matrices (Figure 3.6) were produced by Xingpeng Jiang.

Abstract

Due to the tremendous diversity and abundance of microbes in environmental and host-associated environments, adequate characterization of these samples remains a challenge for microbiologists. In order to increase the depth of sampling for diverse bacterial communities, this thesis research developed a novel method for sequencing and assembly of millions of paired-end reads from the 16S rRNA gene (spanning the V3 region; ~200 nucleotides), using Illumina-based next-generation sequencing. To confirm reproducibility and identify a suitable computational pipeline for data analysis, sequence libraries were prepared in duplicate for both a defined mixture of DNA from known cultured bacterial isolates (>1 million post-assembly sequences) and from an Arctic tundra soil sample (>6 million post-assembly sequences). These Illumina 16S rRNA gene libraries represent a substantial increase in number of sequences over all extant next-generation sequencing approaches (e.g. 454 pyrosequencing); the assembly of paired–end offers a methodological advantage by incorporating an initial quality control step for each 16S rRNA gene sequence. This method incorporates indexed primers to enable the characterization of multiple microbial communities in a single flow cell lane and may be readily modified to target other variable regions or genes.

Soil pH is an important determinant of microbial community composition and diversity, yet few studies have characterized the specific effects of pH on individual bacterial taxa within bacterial communities, both abundant and rare. Composite soil samples were collected over two years from an experimentally maintained pH gradient ranging from 4.5 to 7.5 from the Craibstone Experimental Farm (Craibstone, Scotland). Extracted nucleic acids

were characterized by bacterial and group-specific denaturing gradient gel electrophoresis (DGGE) and were sequenced using the Illumina sequencing method describe above. Both methods demonstrated comparable and reproducible shifts within higher taxonomic bacterial groups (e.g. *Acidobacteria*, *Alphaproteobacteria*, *Verrucomicrobia*, and *Gamma-proteobacteria*) across the pH gradient. In addition, non-negative matrix factorization (NMF) was used for the first time on 16S rRNA gene data to identify positively interacting (i.e. co-occurring) operational taxonomic unit (OTU) clusters (i.e. "components"), with abundances that correlated strongly with pH, and sample year to a lesser extent. The OTUs identified by NMF were visualized within principle coordinate analyses of UniFrac distances and subjected to taxonomic network analysis (SSUnique), which plotted OTU abundance and similarity against established taxonomies. Most pH-dependent OTUs identified here would not have been identified by previous methodologies for microbial community profiling and were unrelated to known lineages.

Methods to limit and reduce carbon emissions are becoming increasingly important for circumventing future impacts of climate change. Biochar is a recalcitrant aromatic-carbon compound formed during pyrolysis in an anoxic environment. The use of lignocellulosic waste material as an input for biochar generation acts as a carbon sink when applied as a soil amendment. Biochar added to soil has been shown to have beneficial effects on crop yield, soil pH, nutrient retention, and fertilizer requirement. However, impacts of biochar applications on soil microbial communities are not well characterized. In order to assess the impact of biochar application on soil microbial communities, two studies were conducted: a multi-year Canadian field trial and a controlled microcosm study. Together, these studies

enabled the assessment of the microbial response to biochar, both with and without the influence of above-ground vegetation, respectively. Field trial samples were collected in 2010, with rhizosphere and bulk soil taken from agricultural plots planted with corn, switchgrass, and soybean, amended with either 0 or 20 t ha^{-1} of dry biochar. The field experiment was also performed on two contrasting soil types: a sandy soil and a loam soil. The microcosm study was conducted over a period of twenty weeks, with biochar added at rates equivalent to 0, 20, 40, and 60 t ha^{-1} to a loam soil in an anoxic incubation system (1 L Mason jar). Nucleic acids were extracted from these soil samples and used as template for bacterial 16S rRNA gene fingerprinting (denaturing gradient gel electrophoresis; DGGE) and amplicon sequencing (101,448,506 assembled sequences generated by Illumina). The resulting fingerprints and PCoA plots based on UniFrac distances indicated that the largest factor governing the microbial community in the field study was soil type, followed by plant type. On the other hand, the corresponding PCoA plots for the microcosm study showed strong separation between biochar-amended samples and controls, in addition to separation corresponding to incubation time. DGGE fingerprints for the microcosm study showed a predominant biochar-associated band. The corresponding sequence in the Illumina libraries classified as an uncultured *Gammaproteobacteria* clone and increased in abundance in biochar-amended samples and was absent from the no-biochar controls. These results indicate that microbial communities detected in the field were controlled primarily by soil type and vegetation cover rather than biochar application, but strong biochar-dependent shifts were observed in the microcosm study.

Acknowledgements

I would like to thank my supervisor, Josh Neufeld, my committee members, Barbara Butler, Trevor Charles, and Gabriel Moreno-Hagelsieb for their helpful input and guidance during this PhD. Thank you to Micheal Lynch, Andre Masella, and Xingpeng Jiang for their excellent bioinformatic assistance and advice and thank you also to the Neufeld lab members, both past and present. For help with field-work and/or sample collection, thank you Graeme Nicol, Rachel Backer, and Ann Balasubramaniam.

I would also like to thank Jennifer Hood, Ann Balasubramaniam, Erin Jones, Jana Tondu, and Robert Garbary for their continued emotional support, perspective and encouragement. Without such wonderful friends, this journey that is graduate school would not have been possible. And finally I would like to thank my family, in particular my parents, Barbara and David Bartram, for their never-ending support.

Table of Contents

List of Figures

List of Tables

Chapter 1
Introduction

1.1 Microbial diversity

1.1.1 Global importance of microorganisms

Microorganisms colonize almost every habitat on the planet, including our bodies (Andersson et al., 2008), soils (Roesch et al., 2007), aquatic environments (del Giorgio and Cole 1998), deep-sea vents operating independently of the sun (Vetriani et al., 2014), and the internal surfaces of rocks (Nyyssonen et al., 2014). In all of these locations, microorganisms catalyze essential processes, of which biogeochemical cycling of carbon and other nutrients is of critical importance (Vogel 2009). Ultimately, Earth's climate, ecosystem health, and the productivities of soil and aquatic environments are dependent on myriad bacterial, archaeal, and eukaryotic microorganisms.

In addition to near ubiquitous distributions, microorganisms are characterized, collectively, by an enormous reservoir of genomic variation, reflected in extensive taxonomic and phylogenetic diversity. As a result of this unparalleled diversity, the discovery of novel enzymes and corresponding coding regions from microorganisms has led to many advances in biotechnology (McMahon et al., 2007; Schloss and Handelsman 2003; Voget et al., 2006), linked to novel enzyme discoveries useful to many industries, including biofuel and bioproduct synthesis, food production, and biomedical applications (Singh and Macdonald 2010). Microorganisms provide humans with additional services that include the breakdown of unwanted anthropogenic substance, such as pesticides (Tyler et al., 2013), treatment of

wastewater (Daims et al., 2006), and important contributions to digestion and nutrition (Dethlefsen et al., 2008).

1.1.2 Measures of diversity

Complex microbial communities, such as those found in soils, represent enormous biodiversity, with estimates ranging from 2,000 to 50,000 unique bacterial species (or "phylotypes") per gram of soil (Roesch et al., 2007; Schloss and Handelsman 2003; Torsvik et al., 1998). Diversity is defined in this thesis as both richness and evenness of the community, where richness describes the overall number of phylotypes in a given sample and evenness represents community abundance distributions, regardless of richness. For example, a community with low evenness might have very high abundances of one phylotype, relative to other phylotypes, whereas an even community may have very similar abundances for many unique phylotypes (Whittaker 1972). With the potential for enormous diversity and extremely small cell sizes, a challenge has been to adequately sample microorganisms for measuring the extent of their individual sample diversity ("alpha diversity"), how the diversity of microorganisms changes across habitat gradients ("beta diversity"), and the combined diversity of all organisms across a landscape ("gamma diversity"; Whittaker 1972). As a result, microbiologists have been limited traditionally by a poor understanding of how microbial communities are structured, and how they contribute to overall ecosystem services (Curtis et al., 2002). Prior to molecular methods, microbial community surveys relied largely on microscopy and the cultivation of microorganisms, which underrepresents diversity. Although having an organism in culture is beneficial in many ways, sometimes irreplaceably so (Rappe and Giovannoni 2003), very few microorganisms are readily isolated by laboratory

cultivation (Amann et al., 1995). As a result, cultivation-independent methods have provided practical alternatives for the study of microbial community composition and diversity, especially for microbial communities that inhabit terrestrial environments.

1.2 Microbial community characterization

1.2.1 The 16S rRNA gene

The ribosomes within all known organisms contain both a small subunit (SSU) and a large subunit (LSU). The SSU (or 30S subunit for Bacteria and Archaea; "S" is for Svedberg units, which is a measure of sedimentation rate) is constructed from the 16S RNA in addition to associated proteins. In addition to ribosomal proteins, the LSU (or 50S subunit for Bacteria and Archaea) consists of the 23S and 5S RNA molecules. Together, the 30S and 50S subunits combine to form the functional 70S ribosome for Bacteria and Archaea, which is smaller than the corresponding ribosome of Eukaryotes (80S).

The 16S rRNA gene codes for a ~1,500 base RNA molecule that has been used traditionally as a biomarker for determining taxonomic and phylogenetic affiliations of microorganisms and for microbial community analysis. This gene, and its 18S homologue, is useful for phylogenetic analysis because of their universal distribution and relatively low frequency of horizontal gene transfer. Low rates of horizontal gene transfer are inferred from correlations between 16S rRNA genes and associated genome similarities within characterized bacteria (Goebel and Stackebrandt 1994).

Because ribosomes serve such essential protein synthesis roles in all living things, the genes that code for the ribosome are highly conserved homologues across all life.

Nonetheless, ribosomal genes possess regions of relatively high and low conservation that reflect the secondary and tertiary structures of the genes within the ribosome itself. The 16S rRNA gene has nine hypervariable regions (V1 to V9), each with sequence and length variability. These variable regions correspond to the hairpin structures in the secondary structure of the rRNA molecule (Neefs et al., 1993). These regions of high sequence variability can reveal important phylogenetic information about species relatedness within and between samples (Pace *et al*., 1986) and, because of this, are often the amplicons of choice in sequencing experiments. This is possible because interspaced between these variable regions are areas of sequence conservation corresponding to secondary structure features of the molecule, and also serving as primer binding sites for polymerase chain reaction (PCR) amplifications. Depending on the degree of conservation, universal or phylum/group specific primers can be synthesized to target desired groups.

Ribosomal genes were essential for establishing a universal phylogeny (Olsen et al., 1986), which revealed the three primary domains of life and the existence of the archaeal phylum (Woese 1987). Prior to the advent of Sanger sequencing, 5S rRNA (~120 nt) was originally used for early phylogenetic work (Olsen et al., 1986). As cloning and sequencing became more commonplace, coupled with newer sequencing methods, the 16S rRNA gene was eventually used (Fox et al., 1980), which was further enabled following the advent of PCR for gene amplification (Lane 1991; Saiki et al., 1985). As of January 14th, 2015, the ribosomal database project (Cole et al., 2014) contains 3,019,928 16S rRNA sequences (in addition to 102,901 fungal 28S rRNA sequences), which can be used in comparisons of unknown 16S rRNA genes from new organisms or uncharacterized environmental samples to

a collection of known sequences, providing putative taxonomic affiliations and phylogenetic information.

Although collected ribosomal genes provide a measure of community composition, analysis of bacterial and archaeal communities with 16S rRNA genes has several limitations. Linking taxonomic information to an organism's functional capabilities are not always straightforward and can be complicated by functional redundancy. In addition, duplications of the rRNA operon commonly exist in many bacteria. Variation in 16S rRNA gene copies ranges between 1 to 15 copies (Acinas et al., 2004). The gene duplications can also exhibit slight to large differences in sequence, complicating analysis and suggesting that horizontal gene transfer may occur and that gene is not always conserved (Doolittle 1999). Copy numbers of rRNA operons reflect an organism's response time to nutrients. With many rRNA gene copies, microorganisms can generate ribosomes rapidly, which can then lead to increased translation rates that equates to an increase in metabolic capacity (Klappenbach et al., 2000). Frequently, however, these differing ribosomal rRNA sequences are within the 97% "species" cutoff commonly employed when processing sequencing data (See Chapter 2; Fig. 2.3A).

Other genes have been proposed as alternate phylogenetic marker genes. Generally, like rRNA genes, these too should be conserved and demonstrate a low frequency of horizontal gene transfer (Neufeld and Mohn 2005a). Genes coding for ribosomal proteins have been proposed as alternative phylogenetic markers (Jolley et al., 2012), as have RNA polymerase genes (Shu and Jiao 2013), and the chaperonin gene *cpn60* (Hill et al., 2002), which also tends to be conserved and is limited to one copy within microbial genomes.

Despite these alternative phylogenetic marker genes, the ribosomal rRNA gene remains the most commonly used marker gene for microbial community analysis and comparative phylogenetics.

1.2.2 Community fingerprinting techniques

Fingerprinting-based methodologies, such as denaturing gradient gel electrophoresis (DGGE; Muyzer et al., 1993) and terminal restriction fragment length polymorphism (T-RFLP; Liu et al., 1997), have been used extensively for characterizing the 16S rRNA gene composition of microbial communities. DGGE generates banding patterns ("fingerprints") from PCR amplicons by differentially separating amplicons of the same length based on characteristics of sequence composition, especially G and C content (Green et al., 2010). Separation is accomplished by electrophoresis at a constant temperature, yet within an increasing concentration of chemical denaturants (i.e. formamide and urea), in a polyacrylamide gel. Band formation occurs when the DNA duplex of PCR amplicons is partially denatured (i.e. "melted"), resulting from the differential bond strengths of A-T and C-G pairs due to the greater number of hydrogen bonds present in C-G base pairings. In contrast, T-RFLP involves labeling of primers with a fluorescent molecule and then treating the resulting PCR products with restriction enzyme digestion. The mixture of digested DNA fragments are resolved on a polyacrylamide gel, each band considered to represent a different phylotype (Liu et al., 1997). T-RFLP has been used in various studies to assess soil microbial diversity (e.g. Doroghazi and Buckley 2008; Dunbar et al., 2000; Fierer and Jackson 2006).

An early cited limitation of DGGE is its inability to detect 16S rRNA gene amplicons (i.e. microbial phylotypes) if individual organisms represent less than 1% of the total

community (Muyzer et al., 1993). In addition, a criticism of T-RFLP has focused on the inability of the method to distinguish between different taxa that generate the same sized fragments, as well as the problem of fluorescence markers having a detection threshold that excludes rare taxa (Blackwood et al., 2007). Despite this valid criticism, Fierer (2007) demonstrates that the utility of T-RFLP lies with it providing a "coarse" representation of the true diversity of a sample. As a result, gel fingerprint-based methods are best suited to rapid and superficial analysis of community composition, identifying relatively large changes in beta diversity across multiple samples, or for the characterization of environments possessing relatively low alpha diversity (Green et al., 2010).

1.2.3 Other methods of microbial community investigation

In addition to fingerprinting methods, other avenues of characterization exist such as fluorescent *in situ* hybridization (FISH) and metagenomics. The FISH method involves hybridizing a nucleotide probe attached to a fluorophore to a glass slide containing fixed cells, with the probe binding to complementary nucleic acid (Amann et al., 2001; Janvier et al., 2003). Instead of single gene sequencing, metagenomics is an approach for gaining information about the genetic and phylogenetic composition, providing information on the functional potential of a microbial community (Blow 2008). For metagenomic analysis, DNA is extracted from an environmental sample without the need for PCR, giving a direct assessment of the genes present. One drawback to metagenomic analyses is that highly diverse habitats such as soil can be difficult to analyze because existing database annotations are unable to adequately describe the sequences obtained from many diverse organisms and their poorly characterized genomes (Ahmed et al., 2008).

1.2.4 Sequencing

In contrast to fingerprinting-based approaches for profiling microbial communities, sequencing of 16S rRNA genes can provide extensive phylogenetic information and detect members of a microbial community members at lower relative abundances than the ~1% detection limit often cited for DGGE. Sequencing-based approaches have been regarded, historically, as being more costly and time consuming (Neufeld and Mohn 2006). However, as sequencing technology has advanced rapidly, the cost per sequence has decreased substantially.

Clone library preparation and Sanger sequencing of near full-length 16S rRNA genes was commonplace before the advent of next-generation sequencing technologies (e.g. pyrosequencing and Illumina). However, these studies were restricted by insufficient sampling depth due to the cost and labour limitations of Sanger-based community profiling of the 16S rRNA gene. For example, one early community analysis examined Amazonian soils in relation to possible deforestation effects and did not identify even two identical sequences among library sizes of 50 sequences per sample (Borneman and Triplett 1997). Similarly, by cloning and sequencing rRNA gene PCR products from a Hawaiian soil, the authors concluded that the diversity exceeded the sequencing effort employed, providing only limited information about overall community structure (Nüsslein and Tiedje 1998). Together, these examples are among many studies characterized by insufficient depth in the sampling of microbial communities, preventing meaningful descriptions and comparisons of microbial community composition (Neufeld et al., 2004).

Recently, microbial ecology methodology has circumvented the cost and labour limitations of traditional clone library sequencing of 16S rRNA genes. Alternative sequencing approaches and the advent of "next-generation sequencing" platforms have increased sequence library sizes greatly and reduced cost and time commitments of microbial community analyses (Prosser et al., 2007). An early alternative approach for sequencing 16S rRNA genes involved sequencing concatamers of short and highly variable 16S rRNA genes. Neufeld et al., (2004) developed serial analysis of ribosomal sequence tags (SARST) to sequence a specific portion of the 16S rRNA gene (V1 region) from soil bacterial communities. This approach involved the concatenation of multiple ribosomal sequence tags (RSTs) prior to cloning and sequencing, producing 10 to 20 RSTs per sequencing reaction, translating into thousands of RSTs at a fraction of the time and costs of traditional clone library sequencing (Neufeld et al., 2004). Others used similar approaches targeting an alternative variable region for accomplishing the same methodological goal (Ashby et al., 2007; Kysela et al., 2005). Although reducing the size of the sequenced fragment reduced taxonomic resolution inferred from each sequenced read (Wang et al., 2007), well-chosen RSTs have high taxonomic resolution. For example, at least 71% of tested RSTs were specific to the genus- or species-level (Neufeld et al., 2004).

The development of next-generation sequencing provided alternative sequencing methods independent of clone libraries, which allows for larger numbers of sequences collected per sample. Pyrosequencing, using 454-based technology, employs the use of oil-in-water amplification of single DNA strands, followed by sequencing by synthesis (Shendure and Ji 2008). Using this technology, Sogin et al., (2006), sequenced V6 variable

regions of samples from multiple deep sea microbial communities. Despite collecting 6,505 to 23,000 16S rRNA gene sequences per sample, none of the species accumulation curves reached an asymptote, indicating that the communities were not sampled to completion. Importantly, this first next-generation analysis revealed a large proportion of relatively rare sequences, which were first attributed to the "rare biosphere". Pyrosequencing was also used initially to profile 16S rRNA genes from four soils, resulting in microbial diversity estimates of 52,000 16S rRNA gene phylotypes per gram of boreal forest soil (Roesch et al., 2007). This unprecedented sampling effort consisted of 149,000 sequences in total (i.e. 26,140 to 53,533 per site). In a subsequent publication from the same research group, the authors calculated that 400,000 to upwards of 1,800,000 sequences per gram of soil would have been required for a complete census of these soil bacterial communities (Fulthorpe et al., 2008).

Illumina is another sequencing technology that is becoming commonplace in microbial ecology studies (Caporaso et al., 2010c; Gloor et al., 2010; Lazarevic et al., 2009) due to its cost effectiveness relative to pyrosequencing (Shendure and Ji 2008). However, before the method reported in this thesis was published (Bartram et al., 2011), Illumina sequence length was restrictive (Caporaso et al., 2010c; Gloor et al., 2010; Hummelen et al., 2010; Lazarevic et al., 2009; Maughan et al., 2012). In addition, prior publications did not fully utilize the error correcting ability of using paired-end sequencing (Caporaso et al., 2010c; Claesson et al., 2010; Lazarevic et al., 2009). However, since publication of the method contained within this thesis, other researchers have corrected for some of these issues (Zhou et al., 2010b).

1.2.5 Bioinformatic approaches and tools

The taxonomic diversity of ribosomal RNA gene libraries and the increased size of sequence datasets presents computational demands. Taxonomic diversity is commonly assessed by a comparison with a reference database and phylogenetic diversity by grouping or clustering sequence reads based on percent similarity (Sogin et al., 2006). Genbank (Benson et al., 2005), which stores all sequenced nucleic acid data, and the Ribosomal Database Project (RDP), which includes ribosomal RNA sequences (Cole et al., 2014), are two examples of such reference sequence databases. Sequences can be compared against these databases using search algorithms, such as BLAST or the RDP classifier. Operational taxonomic units (OTUs) can be generated by clustering sequences with algorithms such as UCLUST (Edgar 2010), CD-HIT (Li and Godzik 2006), or UPARSE (Edgar 2013), beginning with an alignment of all sequences. Alignments are usually generated by identifying regions of similarity with algorithms such as Infernal (Nawrocki et al., 2009), which considers the secondary structure of the rRNA molecule when aligning, NAST and pyNAST (Caporaso et al., 2009; DeSantis et al., 2006). Statistical approaches exist for analyzing the diversity of 16S rRNA gene sequences from environmental samples, including non-parametric estimators, accumulation curves, and the generation of rank abundance curves (Bohannan and Hughes 2003). Non-parametric richness estimators include Chao1 and ACE statistics (Chao 1984; Chao and Yang 1993), and Shannon diversity estimates, which assess dataset richness and evenness.

Beta-diversity characterization methods include UniFrac-based principle coordinate analysis (PCoA) and much less frequently used, non-negative matrix factorization (NMF;

Jiang et al. 2011). UniFrac is a commonly used metric for measuring beta diversity that incorporates phylogenetic information into account when assigning distances (Lozupone et al., 2006; Lozupone et al., 2010). NMF has not been used for 16S rRNA gene data analysis prior to this thesis research, but because of its value for resolving patterns in metagenomic datasets (Jiang et al., 2012; Jiang et al., 2011), it also has strong potential for identifying trends and component taxa in 16S rRNA gene amplicon datasets (Chapter 3).

1.1 The soil environment

1.1.1 What is soil?

Soil can be defined as organic matter combined with minerals from the Earth's crust. Alternately, soil can be thought of as the Earth's surface intertwined with plant roots, an interface where living organisms overlap with inorganic minerals, water, and decaying organic material (Paul 2007). Soils are formed from the breakdown of the underlying parent minerals and rocks via chemical and mechanical weathering, with the addition of early colonizers and subsequent decaying organic matter. Gasses present in pores, water, and dissolved minerals also contribute to the soil environment. The type of soil that forms in any given area is dependent, to varying degrees, on five factors which include the climate, topography or physical features of the region, the parent mineral material, time, and biota (e.g. microbes and vegetation, primarily; Paul 2007).

1.2 Factors affecting microbial communities found in soil

One of the oldest recognized patterns in ecology is that the diversity of plants and animals that are associated with latitude, which has been referred to as the latitudinal biodiversity gradient (Willis and Whittaker 2002). In contrast to the readily observed trends associated with "macroorganisms", identifying the factors that affect soil microorganisms, and describing their diversity, has been extremely challenging due to methodological limitations. Nonetheless, recent studies have leveraged molecular methods, including next generation sequencing, to identify the impacts on soil microbial diversity and composition by factors such as salinity (Lozupone and Knight 2007), substrate availability (Langenheder and Prosser 2008), horizon depth (Zhou et al., 2002), differing land management regimes, and pH (Fierer and Jackson 2006). These factors have all been found to influence microbial diversity to varying degrees. Soil temperature can also affect the composition and rate of activity of soil microorganisms. Because solubility and diffusion of molecules are directly related to temperature, these two processes have substantial influences on soil microbial activity (Paul 2007). Of the factors that impact soil microbial community composition, pH is now recognized as the strongest overall predictor of soil microbial community composition and diversity on a continental scale (Fierer and Jackson 2006).

1.2.1 pH

Soil pH is influenced by mineral composition as well as atmospheric inputs. Protons originating from atmospheric inputs such as rainfall and from organic matter combine with basic material present in soil (e.g. carbonates and aluminosilicates) to establish soil pH. The climate influences temporal pH changes and trends in the soil. For example, soils present in

humid conditions will be subject to exchange of cations from minerals with H^+, deceasing the pH, whereas more arid conditions contribute to the alkalization of soils usually accompanied by sodium (Paul 2007).

Soil pH has a large effect on the relative abundance of cross-kingdom soil decomposers (Rousk et al., 2009). Soil pH can also affect the solubility of inorganic and organic molecules that can directly influence enzyme activity. Pyrosequencing studies demonstrated a positive correlation between pH and diversity, both at a large continental scale (to about pH 8; Lauber et al., 2009) and across smaller, artificially maintained pH specific plots of the same soil type (Rousk et al., 2010). Although influencing overall bacterial diversity, pH differentially impacts specific bacterial groups. For example, groups within the phylum *Acidobacteria* are associated with specific soil pH values (Jones et al., 2009). *Acidobacteria* subgroups 4, 6, 16, and 17 correlated positively with increasing pH, whereas subgroups 1, 2, and 3 correlated negatively with soil pH.

1.3 Biochar

Global climate change, influenced by anthropogenic accumulations of atmospheric CO_2, has become an increasingly urgent concern (Smith et al., 2013; Solomon et al., 2009). Measures to reduce carbon emissions are important for circumventing future impacts of climate change. Fossil fuel use and agricultural practices are two important anthropogenic CO_2 sources (Cole et al., 1997). In contrast to other sources of CO_2, where reduction may be the most effective mitigation practice, there is potential to re-capture carbon back in

agricultural soils. One such approach for mitigating the release of carbon into the atmosphere, and simultaneously providing improved soil fertility, is through the production and storage of "biochar" in the ground (Mao et al., 2012; Woolf et al., 2010).

Biochar is a recalcitrant aromatic material formed by pyrolysis of plant matter under low oxygen or anoxic conditions at low combustion temperatures. This reaction is exothermic due to the release of gaseous byproducts (i.e. oxygen, methane, carbon dioxide and carbon monoxide), which increase the entropy of the reaction and favors continued biochar formation (Antal and Gronli 2003). Low oxygen conditions lead to incomplete combustion of the parent material (Mohan et al., 2006), which optimizes biochar yield and minimizes gaseous oxidation byproducts. Pyrolysis temperatures can range from 250 to 700˚C, with production process durations ranging from hours to days (Novak et al., 2009; Rutherford et al., 2012). Chemically, biochar is similar to graphite. However, unlike the highly ordered structure of layered graphitic sheets, biochar is much more disordered. Biochar can be produced from a wide range of organic carbon inputs, such as nut hulls, poultry waste and wood. The feedstock material, temperature and duration of pyrolysis all have affects on the resultant biochar properties. For example, biochar that results from pyrolysis at a higher temperature generally exhibit a higher pH, larger surface area, and higher ash content (Novak et al., 2009). Following pyrolysis, bio-oil, syngas, and heat are produced in addition to biochar, which can be used as energy sources (Woolf et al., 2010). Biochar differs from ash in carbon content. Whereas biochar is carbon rich, ash contains little carbon and is composed primarily of trace elements. Biochar differs from charcoal in its intended downstream application (Lehmann and Joseph 2009). Biochar is employed for

climate change mitigation, as a soil amendment, as an energy source, and in an effort to manage waste (Lehmann and Joseph 2009). The particular combination of these end-results depends on the type and quality of input biomass.

Biochar has long been used as a soil amendment for low nutrient soils. The first recorded use of biochar for increasing soil fertility was in pre-Columbian South America, originating 8,700 to 500 years before the present. Evidence for this activity is referred to as "terra preta" or "dark earth" (Grossman et al., 2010). These Amazonian Anthrosols exhibit a striking contrast to the surrounding, unmodified and highly weathered Ferralsol (in this case a red clay soil), displaying an increased carbon content and cation exchange capacity, as well as higher levels of nutrients available for plant uptake (Grossman et al., 2010).

The use of lignocellulosic waste material as an input for biochar generation acts as a carbon sink when applied as a soil amendment. The half-life of biochar is estimated to be 10^2 to 10^3 years due to its condensed aromatic structures. This ensures a long residency time in a soil environment where abiotic and microbial oxidation and release of biochar carbon into the atmosphere is minimized (Gonzalez et al., 2005). When added to soils, biochar demonstrates different effects on crop yield, soil pH, nutrient retention, and fertilizer requirements. The impact of biochar on soil pH is of particular interest for unproductive acidic tropical soils. For example, biochar is an economical alternative to lime for soil improvement on the African continent (Bougnom et al., 2010). The impact of biochar on soil pH is dependent on the initial pH of soil in addition to the specific properties and parent material of the biochar (Lehmann et al., 2011).

The impacts of biochar application on a variety of soil types and the microbial communities it influences are still not well characterized. Recently, Takentaini et al. (2013) used pyrosequencing to investigate the impact of biochar on Amazonian Anthrosols and adjacent soils. *Acidobacteria* represented a large proportion of the 16S rRNA gene reads, with acidobacterial subgroups 5 and 6 being more abundant in the biochar-amended soils (Taketani et al., 2013). The authors concluded that due to the stochastic nature of soil formation, the outcome of the addition of biochar was difficult to characterize and predict for future applications.

1.4 Thesis research goals

Even with the advances in methodologies leading to an increase in 16S rRNA genes sampled from different environments, recent studies have indicated that increased sequencing effort is still required. Unanswered questions include: how are microbial communities and their corresponding metabolic processes influenced by biotic and abiotic factors? How do these communities influence the functioning of ecosystems? What effects do anthropogenic disturbance, such as climate change, have on the diversity and function of microbial communities? Developing and applying novel molecular methods for vastly increased profiling of microbial communities will help improve our understanding of the microbial diversity and composition present in complex ecosystems such as soil.

The broad goal of this research was to better understand the factors affecting microbial community composition and diversity in terrestrial environments. A critical step towards this goal was the development of a high-throughput sequencing method to quantify

microbial diversity in complex soil communities. This research investigated the effect of external chemical gradients (such as pH gradients and biochar application) on bacterial community composition and diversity, by using both high-throughput sequencing and fingerprinting approaches.

Chapter 2*

Generation of multi-million 16S rRNA gene libraries from complex microbial communities by assembling paired-end Illumina reads

2.1 Introduction

The composition, organization and spatial distribution of environmental microbial communities are still poorly understood. Enormous progress in method development has begun to enable the study of alpha, beta, and gamma diversity, but a substantial limitation remains: the coverage of most sequencing methods remains insufficient to analyze single samples comprehensively or conduct field-scale comparisons of the microbial diversity in most environments. Methodology is still required to provide (a) high sample throughput, (b) information on both the microbial species (or phylotypes) present at both high and low relative abundance, and (c) affordability for the average research laboratory. Although comprehensive metagenomic analysis could eventually be used for microbial community profiling (sampling both abundant and rare populations), this is not yet feasible for most environmental samples due to enormous computational and sequencing limitations. Instead, an alternative community profiling approach involves surveying distributions of the small

* A version of this chapter was previously published as "Bartram, A.K., Lynch, M.D.J., Stearns, J.C., Moreno-Hagelsieb, G., and Neufeld, J.D. 2011. Generation of multi-million 16S rRNA gene libraries from complex microbial communities by assembling paired-end Illumina reads. Appl. Environ. Microbiol. **77**: 3846-3852."

subunit ribosomal RNA (rRNA) gene due to its ubiquity across all domains of life (16S rRNA in Bacteria and Archaea; 18S rRNA in Eukarya; (Olsen et al., 1986)). Additionally, the 16S rRNA gene provides valuable phylogenetic information (Curtis et al., 2006) for comparison to database collections. For many years, the use of Sanger sequencing for collected 16S rRNA genes from environmental samples has revealed that sample sizes, and thus coverage, afforded by Sanger sequencing have been insufficient to adequately describe and compare microbial communities (Curtis et al., 2006; Neufeld and Mohn 2006). The advent of serial analysis of ribosomal sequence tags (SARST; (Kysela et al., 2005; Neufeld et al., 2004; Yu et al., 2006)) and 454 pyrosequencing provided a major advance by enabling the collection of thousands of sequences from multiple samples. These approaches have provided a new window into the diversity and composition of microbial communities (Huber et al., 2007; Neufeld and Mohn 2005b; Sogin et al., 2006), increased sample throughput using indexing (Andersson et al., 2008; Hamady et al., 2008; Lauber et al., 2009), and sparked interest in elucidating the members of the rare biosphere, which are microorganisms that exist at low relative abundance (Neufeld et al., 2008; Pedros-Alio 2007; Sogin et al., 2006). To further reduce the costs of sequencing, the Illumina platform has recently been used to generate datasets of unprecedented size (Caporaso et al., 2010c; Gloor et al., 2010; Lazarevic et al., 2009) that surpass 454 pyrosequencing by over an order of magnitude in sequences per unit cost (Shendure and Ji 2008). Initial Illumina-based methods for sequencing 16S rRNA genes have been limited by ≤101 base sequence reads (Caporaso et al., 2010c; Gloor et al., 2010; Hummelen et al., 2010; Lazarevic et al., 2009; Zhou et al., 2010b) and/or an inability to leverage the paired-end approach that would allow for assembly

of reads and reduced sequencing errors (Caporaso et al., 2010c; Claesson et al., 2010; Lazarevic et al., 2009).

Here, a novel application-ready method is presented for generating multi-million sequence datasets at a fraction of the cost of Sanger or 454 pyrosequencing. Without factoring sample preparation costs, Illumina is currently ~50X and ~12000X less expensive than pyrosequencing (i.e. 454) and Sanger sequencing per sequenced megabase, respectively (Sergio Pereira personal communication; The Centre for Applied Genomics, Toronto, Canada). This method uses the paired-end Illumina sequencing platforms (i.e. GAIIx Genome Analyzer, Hiseq 2000, and MiSeq Genome Analyzer) to assemble ~200 base hypervariable region (V3) amplicons with individual forward and reverse read lengths of 125 nucleotides each. We demonstrate with replicate defined community and Arctic tundra libraries that 16S rRNA gene sequencing with the Illumina sequencing platform enables rapid, affordable, reproducible, and comprehensive assessments and comparisons of the taxonomic diversity present in complex microbial communities, and provides unprecedented access to organisms present at low relative abundance.

2.2 Materials and Methods

2.2.1 Sample collection and DNA isolation

A composite Arctic tundra soil sample was prepared from a pristine site in Alert, Nunavut, Canada. This soil sample was collected and used previously for analysis of 16S rRNA gene sequence tag data using the SARST technique (Neufeld et al., 2004). For a defined community, six bacterial strains were chosen as controls: *Escherichia coli* (ATCC

11303), *Pseudomonas aeruginosa* (ATCC 10145), *Bacillus subtilis* (ATCC 6633), *Flexibacter canadensis* (ATCC 29591), *Methylococcus capsulatus* str. Bath (ATCC 33009), and *Paracoccus denitrificans* (ATCC 17741). These organisms were chosen to provide wide coverage of genera and rRNA operon copy numbers. Genomic DNA was extracted from soil and log-phase bacterial cultures using the FastDNA spin kit for soil (MP Biomedicals, USA) according to the manufacturer's instructions. Soil DNA was extracted in triplicate and the extracts were subsequently pooled. Ten nanograms of each pure culture template DNA was combined prior to PCR in order to eliminate possible bias associated with DNA extraction.

2.2.2 Illumina library generation.

The hypervariable region 3 (V3) of the 16S rRNA gene was amplified using modified 341F and 518R primers ((Muyzer et al., 1993); Table A-1). In addition to V3–specific priming regions, these primers are complementary to Illumina forward, reverse and multiplexing sequencing primers (with the reverse primer also containing a six base-pair index, allowing for multiplexing). All custom primers were synthesized and purified by polyacrylamide gel electrophoresis (PAGE; IDT, Coralville, IA). Three PCR amplifications were carried out for each sample in 50-µl volumes. Each reaction contained 25 pmoles of each primer, 200 µM of each dNTP, 1.5 mM $MgCl_2$, and 1 U Phusion *Taq* (Finnzyme, Finland). The PCR conditions involved an initial denaturation step at 95˚C for 5 minutes followed by 20 cycles of 95˚C for 1 minute, 50˚C for 1 minute and 72˚C for 1 minute, and ending with an extension step at 72˚C for 7 minutes in a DNA Engine thermocycler (Bio-Rad, Mississauga, ON). Following separation of products from primers and primer dimers by electrophoresis on a 2% agarose gel, PCR products of the correct size were recovered using

the QIAquick Gel Extraction Kit following manufactures instructions (Qiagen, Mississauga, ON). For each library, triplicate soil PCR products with unique indexes were mixed in equal ng quantities, quantified on a NanoDrop ND2000 (Thermo Scientific, Wilmington, DE) and sent to Illumina (Hayward, CA) for 125-nucleotide paired-end multiplex sequencing. The Alert DNA was included in a greater proportion than the defined community (approximately 20:1). Together, the Alert libraries accounted for approximately 75% of the total DNA sent for sequencing in a single lane; other samples unrelated to this study occupied the balance (~25%) of the template mixture. The quality and concentration of the purified library was determined by Agilent Bioanalyzer analysis. The library was clonally amplified on a cluster generation station using Illumina Version 4 cluster generation reagents to achieve a target density of approximately 150,000 clusters per tile in a single channel of a flowcell. The resulting library was then sequenced on a GAIIx Genome Analyzer using Illumina Version 4.0 sequencing reagents, generating paired reads of 2x125 bases. After sequencing, image analysis, base calling and error estimation were performed using Illumina Analysis Pipeline (version 2.6).

2.2.3 Clone libraries.

Either soil or pure culture genomic DNA was used as template with primers 27f and 1492r (Lane 1991) targeting the full-length bacterial 16S rRNA gene. The PCR amplifications were performed in 25-µl volumes with the concentration of reagents and reaction conditions as described for Illumina library generation, with the exception of the extension step, which was extended to 1.5 minutes to accommodate the longer amplicon. Reaction products were cloned into the TOPO vector (Invitrogen, Burlington, ON) according

to manufacturer's instructions. Ninety-five positive clones were selected from each library (either soil or pure culture library) and sequenced with Sanger technology (Beckman Coulter Genomic Services, Danvers, MA).

2.2.4 Initial quality filtering.

Using a custom algorithm (PANDAseq, see supplemental online material), Illumina reads were binned according to index sequence. Overlapping regions within paired-end reads were then aligned to generate "contigs". If a mismatch was discovered, the paired-end sequences involved in the assembly were discarded. All sequences with ambiguous base calls were also discarded.

2.2.5 Bioinformatic analysis.

All sequences (Illumina and Sanger-based) were assigned taxonomic affiliations based on a naïve Bayesian classification (RDP classifier; (Wang et al., 2007)) with an assignment cutoff used of 0.5. Additionally, assembled contigs and Sanger clone library sequences were used as input for modified single-linkage clustering using CD-HIT (Li and Godzik 2006). Good's coverage (Good 1953) was calculated for each of the resulting libraries to estimate the sequence coverage of the composite Alert library (AT). All Illumina sequence data from this study were submitted to the NCBI Sequence Read Archive (SRA) under the accession number SRA024100. Sanger-sequences for the Alert and Control libraries were submitted to Genbank under accession numbers JF508183-JF508359.

2.3 Results

2.3.1 Development of Illumina for 16S rRNA gene sequence analysis.

The V3 region of the 16S rRNA gene was selected for this method because of its taxonomic resolution (Huse et al., 2008) conserved flanking regions (Muyzer et al., 1993) and length (Gloor et al., 2010; ~170-190 nucleotides; Fig. 2.1A), which is compatible with paired-end 125-base read assembly (Fig. 2.1B). Complete variable-region assembly, by virtue of overlapping 3' end sequences, reduces sequencing errors and generates datasets that are compatible with established computational analysis pipelines (e.g. QIIME;Caporaso et al., 2010b). Custom primers (Table A-1) contain regions specific to the Illumina flow cell, unique indexing to allow for multiplexing of samples, and regions complementary to the conserved portions of the bacterial 16S rRNA gene flanking the V3 region. Additional error-correcting indexes (indexes 13 to 84) were designed using the Barcrawl software package (Frank 2009). The bacteria-specific primers are identical to those used for the initial application of denaturing gradient gel electrophoresis (DGGE) for microbial community analysis (Muyzer et al., 1993). The use of a single low-cycle-number PCR step and subsequent gel purification (Fig. 2.1A) greatly decreases hands-on library preparation time compared to previous sequencing approaches. To validate this method, we analyzed a defined mixture of genomic DNA from six microorganisms as a control library (C) and an Arctic tundra soil from Alert in Nunavut, Canada (AT), which was previously analyzed by SARST (Neufeld and Mohn 2005b). Technical replicates of each sample (C1/C2 and AT1/AT2) were performed to confirm the reproducibility of this technique. Paired-end reads were assembled by aligning the 3' ends of forward and reverse reads. This assembly step

provided additional quality control in the lower quality 3' regions of each read (Fig. 2.2), given that Phred scores are additive in the overlapping region. The average assembly overlap was 66 ± 11 bases and the average post-assembly sequence length of our libraries was 150 ± 11 (without sequenced primers). This overlap resulted in 2-fold coverage across a substantial portion of each sequence in our libraries. Paired-end reads that did not assemble as contigs were discarded, as they possessed sequencing errors (presenting as mismatches between the complementary ends of the two reads). This greatly decreased the number of artifactual sequences used in downstream analyses, with almost 50% of sequences omitted from subsequent analysis for replicate control and Alert tundra libraries (Table 2.1).

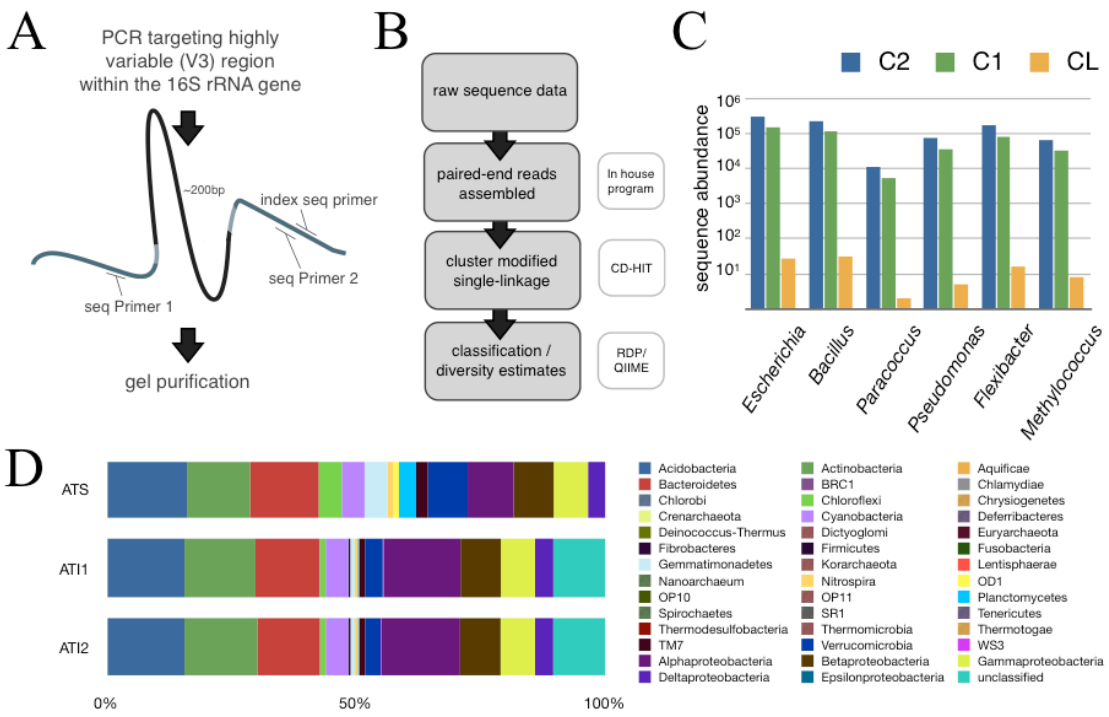

Figure 2.1 Overview of the Illumina 16S rRNA gene sequencing method and generated library data. (A) The schema indicates a PCR (20 cycles) and gel purification of ~330-base PCR products, including the conserved 16S rRNA gene primer-binding region. (B) Informatics pipeline for generating clusters and taxonomic affiliations. (C) Resulting taxonomic affiliations for the replicate control libraries (C1 and C2) and the Sanger sequencing-based library (CL). (D) Taxonomic affiliations for the Alert tundra duplicate libraries (AT1 and AT2) and the Sanger sequencing-based library (ATS).

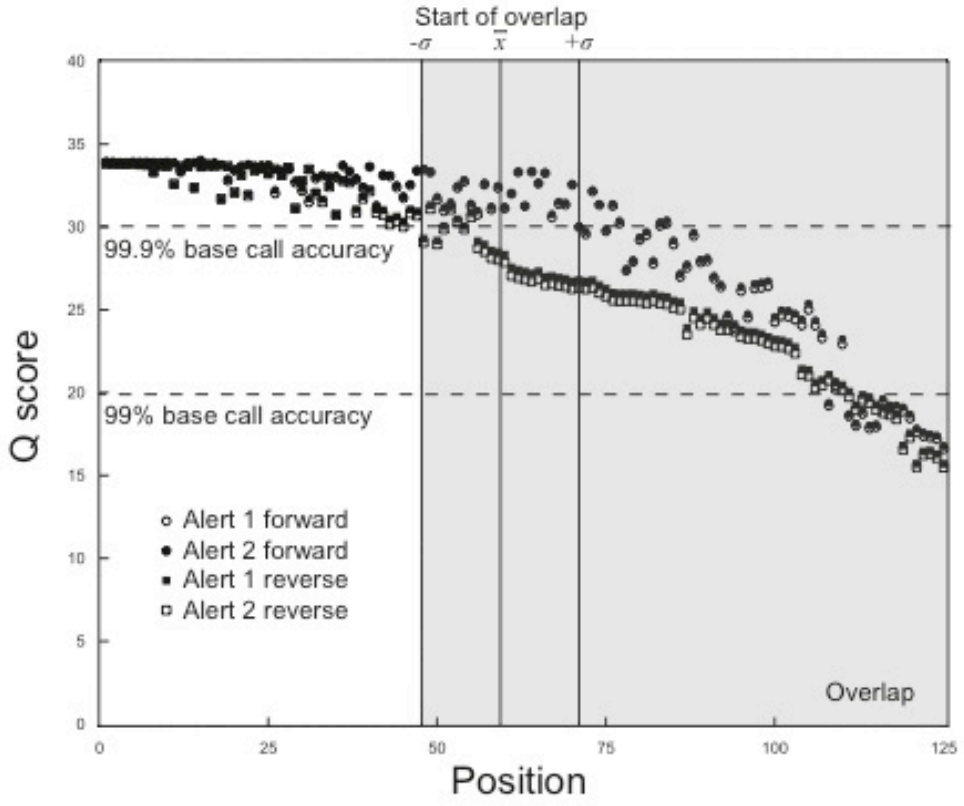

Figure 2.2 Quality (*Q*) scores for all 125-base sequence reads. The *Q* score is an integer mapping of *P*, the probability that the corresponding base call is incorrect, with higher *Q* scores indicating lower error rates. The magnitude of sequence overlap for each assembled read was characterized, and the mean (x) and standard deviation ($\pm\sigma$) were plotted relative to sequence length. The region of potential read overlap as presented does not explicitly calculate the additive *Q* score at each position, as the range of overlap varied due to the large range of V3 lengths.

Table 2.1 Counts of paired-end rRNA gene sequences obtained from the Illumina flow cell (preassembly) and following assembly (postassembly) for the replicate libraries included in this study

Library	Pre-assembly	Post-assembly	Remaining (%)
Alert tundra 1 (AT1)	7,570,249	4,073,963	53.8
Alert tundra 2 (AT2)	4,371,453	2,396,331	54.8
Control 1 (C1)	716,366	464,045	64.8
Control 2 (C2)	1,350,602	842,585	62.4

2.3.2 Defined community clustering and error rates.

To generate taxonomic profiles of the samples included in this study, the assembled sequence data were assigned to taxonomic groups using the naïve Bayesian classifier v.2.1 (Wang et al., 2007) from the Ribosomal Database Project (RDP; Cole et al., 2009). A six-organism defined community was constructed for control purposes by mixing equal proportions of extracted genomic DNA from six bacterial species for the generation of both Illumina and Sanger libraries. The resulting read counts generated by the Illumina method were at least four orders of magnitude higher than the counts of the corresponding sequences generated using a clone library (Fig. 2.1C); the cost of generating the Illumina C1/2 replicate libraries (>1 million sequences total) and the Sanger control library (CL; 95 sequences) were roughly equivalent at the time of sequencing.

Huse and coworkers (Huse et al., 2010) reported that a single-linkage preclustering step

followed by average linkage clustering at 3% gave a more accurate OTU characterization in pyrosequencing datasets and minimally affected the presence and distribution of microbial taxa. Given the large size of our Illumina libraries, linkage clustering would be too computationally intensive. Instead, we used CD-HIT (Li and Godzik 2006) to cluster our control and Arctic tundra datasets at an equivalent 97% sequence identity (Fig. 2.3A). Applying such clustering to C1 and C2 libraries revealed V3-region sequences from the six microorganisms well above background noise (Fig. 2.3A). Clustering the same C1 and C2 libraries at 95% identity had a minimal effect on further reducing low-abundance sequences (data not shown). The 97% clustering step on the assembled control library sequences provided a measure of the total effect of sequencing and PCR errors on the resulting libraries. Clustering of control libraries at 97% identity increased the counts of sequences binned within expected phylotypes by 18.4%, suggesting that approximately one in every five ~200-base sequences (including 16S rRNA gene primers) contained at least one error (~1% error including PCR error). For comparison, the error rate of Sanger sequencing can be as low as 0.001% (excluding PCR errors) with raw error for pyrosequencing (excluding PCR errors) ranging from 1 to 1.5% (Shendure and Ji 2008). However, these single nucleotide errors had little effect on classification of the sequences due to the clustering step. Additionally, sequences were detected in the Illumina libraries (C1/C2) that were not seen in the Sanger libraries (CL), which did not cluster within expected V3 regions of the defined communities. These errors did not appear to be caused by PCR error or chimeras because they were confidently affiliated with 16S rRNA gene sequences from known organisms (Table 2.3A). Instead, these sequences likely resulted from the co-extraction of DNA from the bacterial

growth medium or in low-level contamination of reagents used for PCR, an effect also observed in a recent pyrosequencing study (Huse et al., 2008). If associated with bacterial growth media, these contaminating sequences would not affect results obtained from environmental samples.

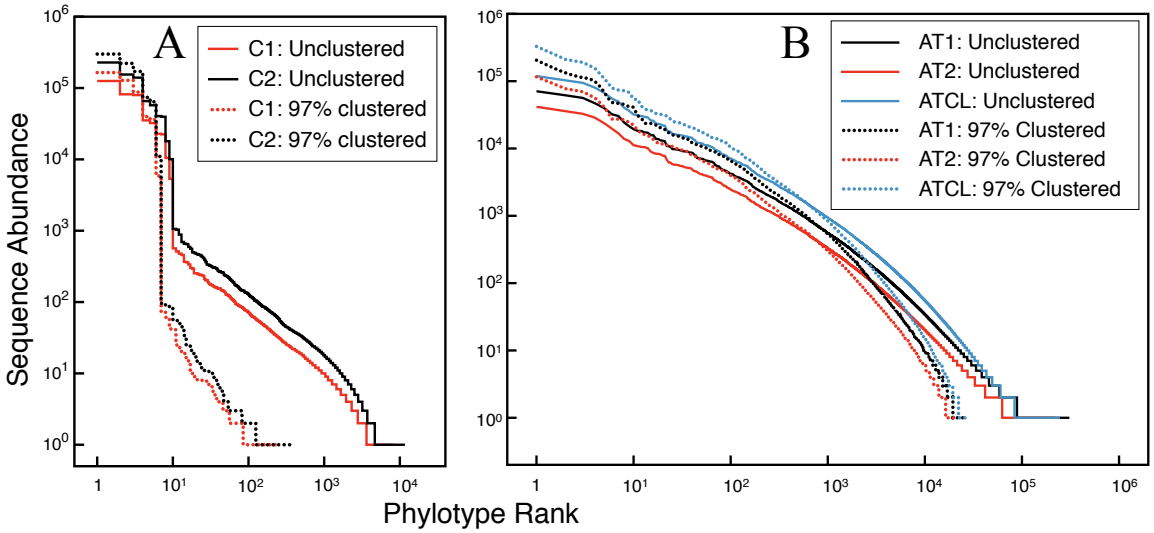

Figure 2.3 Rank-abundance curves for duplicate control libraries (A) and Alert Arctic tundra libraries (B). The data shown are the raw data and also the data clustered using CD-HIT at a cutoff of 97%. Note that the Alert Illumina library was considered as separate replicates (AT1 and AT2) and also as a composite library (ATCL), which represents the combined replicates.

2.3.3 Clustering and characterization of Arctic tundra libraries.

The duplicate Arctic tundra libraries displayed a high degree of similarity based on a comparison of phyla representation to one another (AT1 to AT2; r=0.999) and to a small Sanger-sequenced Arctic tundra clone library (ATS; r=0.950 Fig. 2.1D). Representational differences between the two sequencing approaches is likely due to primer bias, because different primers were used for the construction of each library. The similarity of the V3 sequencing data and the Sanger-based clone library was much higher by overall phyla distribution than between these libraries and additional datasets generated from this same sample using V1-region sequencing (17-55 bases in length) with either SARST (Neufeld and Mohn 2005b) or Illumina-based approaches (Fig. 2.4). Clustering at 97% similarity reduced the proportion of singleton sequences from 5.9% (unclustered) to 0.17% (clustered) of all sequences in the combined AT libraries (Fig. 2.3B) and indicated that high-abundance phylotypes increased disproportionately to low-abundance phylotypes when AT1 was combined with AT2 to form ATCL. Calculated Good's nonparametric coverage estimates for the combined AT Illumina dataset increased from 0.962 to 0.996 for unclustered and clustered libraries, respectively. In contrast, Good's coverage for full length 16S rRNA gene Sanger-based clone library clustered at 97% identity (87 clones) was only 0.207. Good's coverage was also used to assess the effect of library size on coverage, with increasing subsamples of the combined AT dataset. The Good's coverage estimates were >0.95 with >1 million sequences sampled. Additionally, once millions of sequences were sampled, Chao1 richness estimates began to reach an asymptote (Fig. 2.8A). This illustrates that multi-million sequence libraries (generated using Illumina sequencing method) generate high estimates of

completeness of sampling (Fig. 2.5). Replicate Illumina sequence libraries for the Alert tundra DNA sample (AT1/2) were highly similar, with the majority of sequences (99.57%) corresponding to clusters detected in both replicates (Fig. 2.7 inset) and a high Bray-Curtis similarity value (0.96), especially when the 50-most abundant phylotypes were considered (0.99; Fig. 2.8B). Rank-abundance curves for the most abundant phylotypes were nearly identical in distribution (Fig. 2.8B). Although there were several clusters unique to one of the replicate libraries, these were largely composed of clusters represented by a single sequence (Fig. 2.7).

Taxonomic classification of 97% sequence identity clusters demonstrated a distinct taxonomic shift when comparing predominant to low relative abundance clusters or ranks (Fig. 2.6). The ten most abundant ranks accounted for 20.6% of all sequences and belonged to the *Acidobacteria*, *Actinobacteria*, *Bacteroidetes*, *Cyanobacteria* and *Proteobacteria* (Tables A-3, A-4 and A-5). Except for the *Cyanobacteria*, which were largely absent in the lower ranks and singletons, these phyla remained predominant throughout the lower abundance ranks. There was an increase in the number of phyla present in low abundance ranks, with a maximum of 28 phyla represented by the 10001-doubleton abundance rank. There was also a notable increase in the proportion of *Verrucomicrobia* in mid-range abundant ranks (11-100 and 101-1000 most abundant), which were absent in the high abundance ranks (1-10). Furthermore, sequences affiliated with TM7 were only predominant in low-abundance ranks. The proportion of sequences assigned as unclassified (i.e. weakly classified or not classified at all) increased from absent in high abundance ranks (1-10) to approximately 25% of all rare sequence cluster ranks (Table A-3). Future work will aim to

separate errors from genuine diversity, confirming that low abundance sequences are not simply accumulated artifacts from increased sequencing intensity, as suggested by Kunin et al., (2010).

Patterns of rank-specific taxonomic distributions observed in phyla were also present in Class and Order classifications (Fig. 2.6, Tables A-4 and A-5). In each case, the abundant clusters maintained their predominance in lower-abundance ranks (with the notable exception of cyanobacterial sequences), with taxonomic diversity increasing in lower-abundance ranks. Notably, in Class and Order classifications, the presence of clusters labeled unclassified increased incrementally from <5% (1-10 ranks) to ~50% (singletons) for both classifications. The increase of unclassified sequences was both larger for low-abundance clusters within a taxonomic level and increased with depth of classification (i.e. the majority of sequences were successfully classified to Phylum, while even some high-abundant sequences were not successfully classified to the Ordinal level).

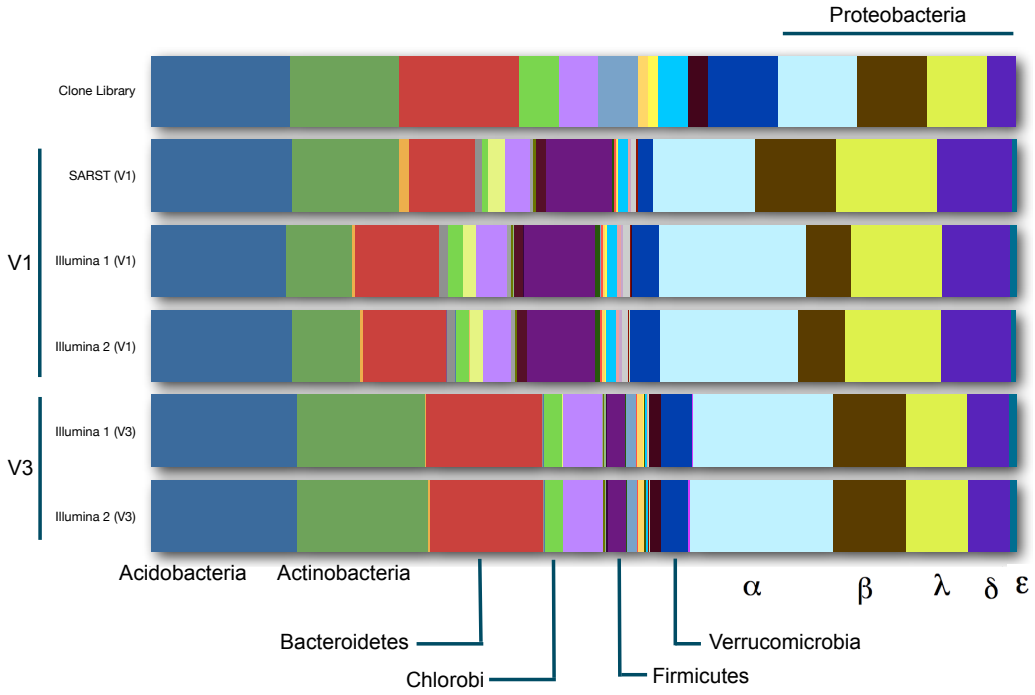

Figure 2.4 Comparison of phyla distributions for the Arctic tundra libraries, using clone library analysis, SARST (previously published), Illumina (both V1 and V3 regions; processed as duplicate libraries). The V1 region dataset represents unpublished data from a previous iteration of this methodology.

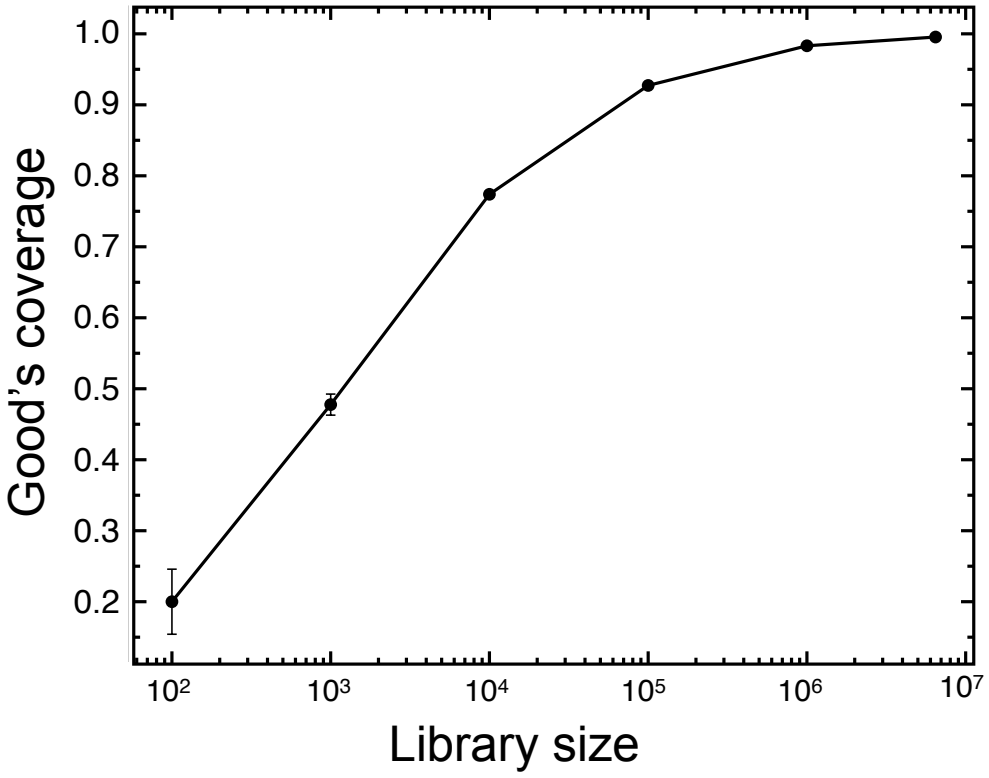

Figure 2.5 Effect of library size on phylotype coverage. Randomly subsampled libraries were drawn in triplicate from combined AT libraries and used to calculate Good's coverage estimates. Averages for triplicates were plotted with standard deviations.

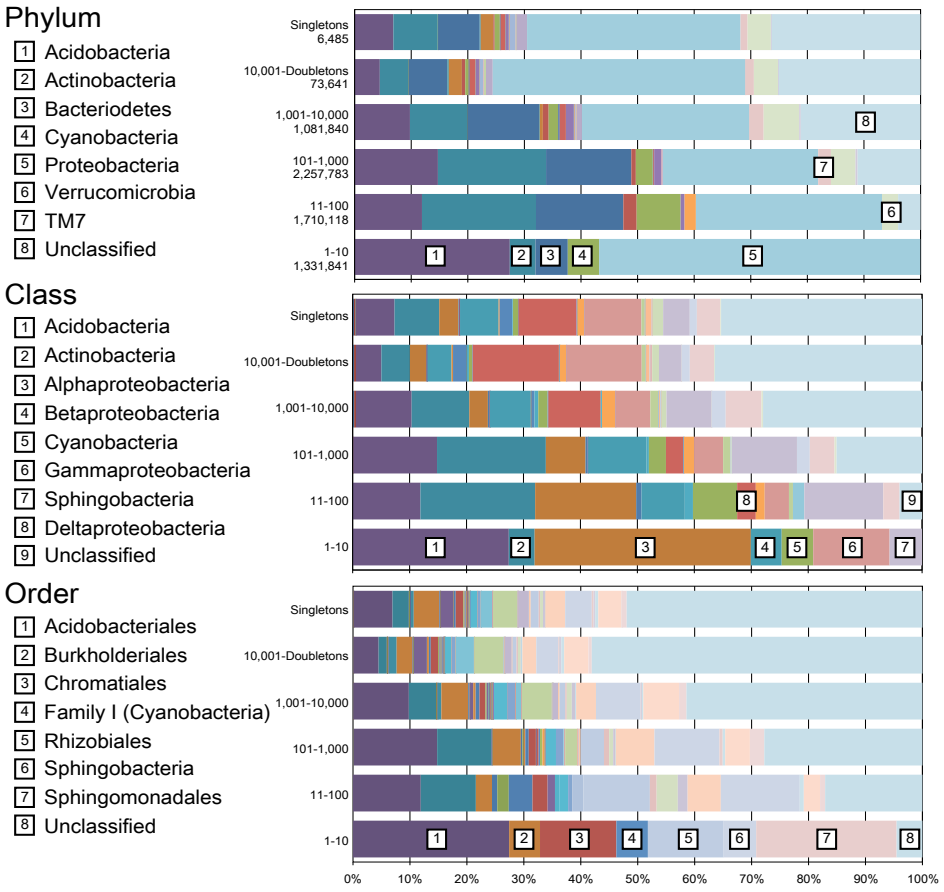

Figure 2.6 Taxonomic affiliations at the levels of phylum, class, and order for consecutive abundance ranks of sequence data clustered at 97% with CD-HIT. Predominant taxa are represented in the bottom row, and singletons are at the top for each taxonomic level. Full details of RDP affiliations are summarized in Tables A-1, A-2, and A-5 in the supplemental material.

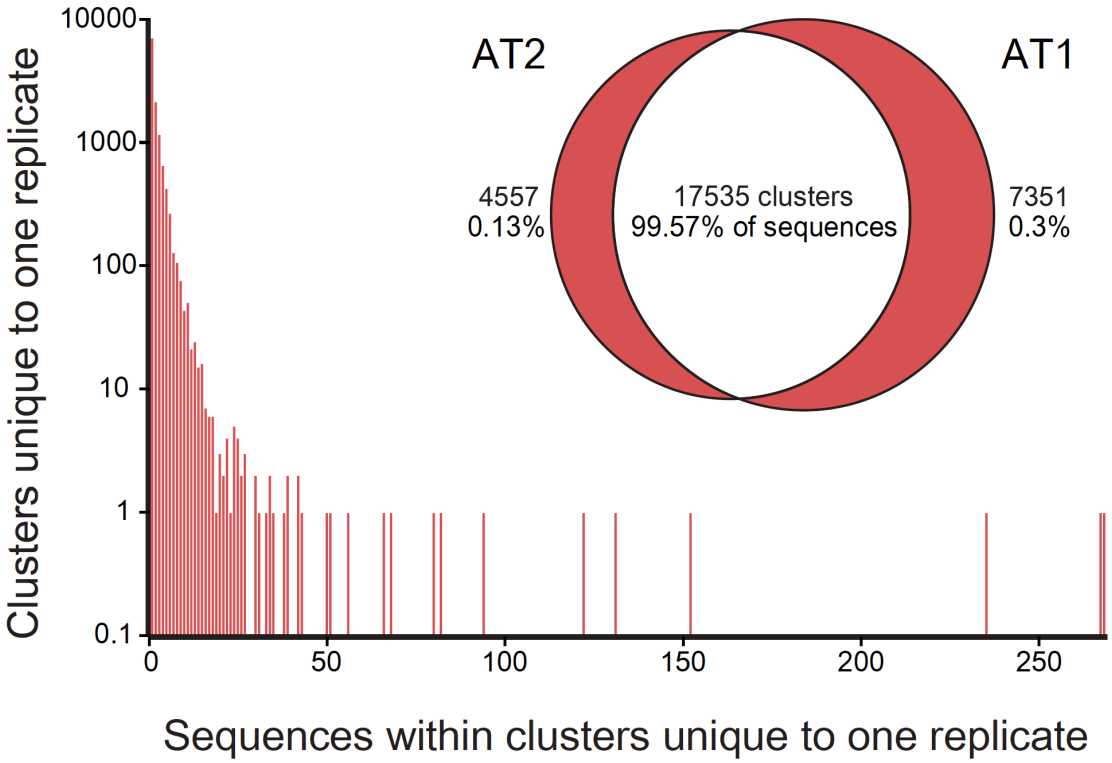

Figure 2.7 Plot of V3-region clusters (phylotypes), which were associated exclusively with either AT1 or AT2. The results demonstrate that the vast majority of clusters associated with one of the replicates were found in singletons and other low-abundance ranks. Inset: Venn diagram of the number (and percent) of clusters associated with either replicate, or with both replicates.

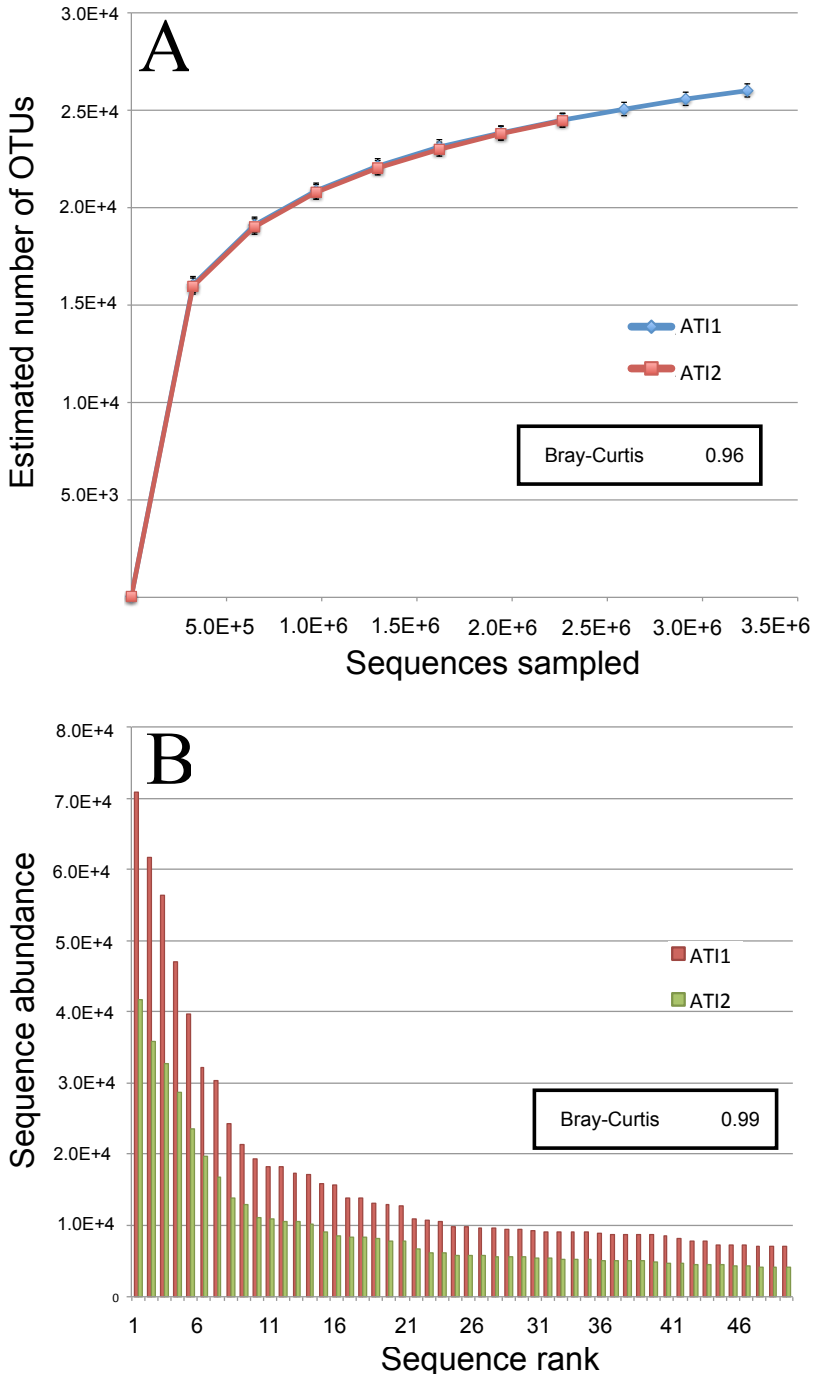

Figure 2.8 Diversity estimates. (A) Chao1 richness estimate of Alert tundra Illumina libraries. Inset: Bray-Curtis similarity metric of proportional abundance of the Alert tundra library replicates. (B) Top 50-ranked most abundant sequences. Inset: Bray-Curtis similarity metric calculated for top 50-ranked sequences (proportional values).

2.4 Discussion

Here we demonstrate improvements in both sampling depth and sequence quality using an inexpensive and rapid sequencing methodology. An advantage of this technique over current high-throughput methods is the assembly of paired-end reads that greatly reduces the number of erroneous sequences included in downstream analyses. Importantly, as the read lengths for the Illumina platform increase (during this study they were ~125 bases) so too will the quality of the libraries generated with this technique. Additionally, the use of index sequences enables many samples to be sequenced in parallel. We have tested 24 indexed primers in our laboratory (data not shown) and additional index sequences have been provided that can further increase sample throughput (Table A-1). Further improvements to this method can be introduced, such as the addition of a highly diverse series of bases adjacent to the forward sequencing primer-binding area (Table A-1; data not shown). This addition improves Illumina base-calling because the algorithm identifies clusters optimally on the flow cell when maximum nucleotide diversity is present across the first four bases sequenced in the forward read. In addition, the long oligonucleotide primers used here were purified commercially with polyacrylamide gel electrophoresis (PAGE) for an additional cost (IDT, Coralville IA). Future research will determine if standard desalting of primers will be sufficient to generate Illumina datasets, which would reduce the start-up cost for this new technology.

With the increase in recovered sequences, there is a corresponding increase in artifact sequences. The capacity of the Illumina platform to generate enormous datasets is undoubtedly an advantage; however, if low abundance phenotype discovery and accurate

measurements of alpha diversity are desired, errors must be effectively managed. Otherwise, community characterization is only useful at a coarse level. In this study, assembly was accomplished by the use of overlapping paired-end reads, and a modified single-linkage clustering protocol was applied at 97% sequence identity. Future work will identify effective clustering algorithms that adequately reduce datasets to the expected phylotype diversity, as shown recently for 454 pyrosequencing data (Huse et al., 2010), and which would be scalable to sequence libraries possessing many millions of sequences and hundreds (or thousands) of samples. Additionally, problems resulting from the sensitivity of the technology (e.g. sequencing of low-abundance sequence contamination in laboratory growth media) would be bypassed by multiplexing PCR amplifications directly from environmental samples as outlined in this protocol.

Regardless of sequencing artifacts, advances in sequencing technologies are paralleled by an increased magnitude of phylotype diversity surveyed from microbial communities. Although a small number of sequences may be sufficient to detect underlying patterns differentiating highly divergent communities (Kuczynski et al., 2010), larger datasets are required to identify more subtle responses to environmental factors among less predominant populations and increased sequence coverage of the rare biosphere (Huse et al., 2010; Sogin et al., 2006). Rare microbial taxa likely represent microorganisms (a) adapted to life at low relative abundance, (b) that have not been discovered previously and (c) possessing abundance distributions with important correlates to measured physicochemical parameters. In this study, the Illumina sequencing platform provided access to low-abundance phylotypes from soil with coverage (Fig. 2.5), and combined library sizes greater than those reported

previously (Caporaso et al., 2010c; Claesson et al., 2010; Rousk et al., 2010). The main limitation of recent iterations of the Illumina platform has been the reduced taxonomic resolution of short sequence reads (Caporaso et al., 2010c; Gloor et al., 2010; Lazarevic et al., 2009). With the introduction of 125-base paired-end reads reported here, this sequencing methodology can now span the taxonomically informative V3 variable region of the 16S rRNA gene and will soon generate twofold coverage of complete PCR amplicons as sequence length continues to increase. Note that the V3 region chosen here was selected because the primers used are the same as those used for DGGE of bacterial communities (Muyzer et al., 1993) and that this region is longer (~170-190 bases) than the V6 region, which was sequenced elsewhere (~105-120 bases; (Gloor et al., 2010)). Although base-calling accuracy decreases markedly toward the 3' end, the sequence read overlap of 66 ± 11 nucleotides (ATCL library) greatly increased data quality in this region (Fig. 2.2). The primers and adaptors are modular; this sequencing methodology can readily be modified to target other genes or regions of interest. This versatile, affordable and powerful methodology greatly increases the depth at which low-abundance organisms can now be probed, noted by the high Good's coverage estimates (Fig. 2.5), high similarity between replicates (Fig. 2.8) and the number of unclassified or unique taxa in low abundance groups (Fig. 2.6), suggesting that we are now able to comprehensively and reproducibly characterize and compare abundant and rare populations across multiple samples derived from complex microbial communities.

Chapter 3[†]

Exploring links between pH and bacterial community composition in soils from the Craibstone experimental farm

3.1 Introduction

Soil microbial communities are important contributors to biogeochemical processes and are characterized by high taxonomic and metabolic diversity (Prosser et al., 2007). Despite their global importance, a lack of empirical knowledge remains regarding the factors that affect soil microbial community composition. Recently, next-generation sequencing technologies (e.g. pyrosequencing and Illumina) have helped identify factors that influence soil microbial diversity, ranging from salinity (Lozupone and Knight 2007), metal contamination (Gans et al., 2005), resource availability (Langenheder and Prosser 2008), depth and water availability (Eilers et al., 2012; Zhou et al., 2002). Importantly, soil pH represents the strongest known predictor of microbial community composition and diversity in surface soils, with an R^2 value of 0.70 when phylotype diversity and pH were examined (Fierer *et al*., 2006). Single-gene sequence data sets generated by pyrosequencing demonstrate a positive correlation between alkalinity (measured by pH) and diversity at a continental scale (below c. pH 8; (Lauber et al., 2009)), in Arctic tundra (Chu et al., 2010) and across an experimental pH gradient within a single soil type (Rousk et al., 2010).

[†] A version of this chapter was previously published as "Bartram, A.K., Jiang, X., Lynch, M.D.J., Masella, A.P., Nicol, G.W., Dushoff, J., Neufeld, J.D. Exploring links between pH and bacterial community composition in soils from the Craibstone experimental farm. FEMS microbiol. Ecol. **87**. 403-415."

The experimental plot results demonstrated that although bacterial and fungal abundance responded variably to soil pH, both bacterial and fungal diversity increased with increasing pH (Rousk et al., 2010). This study also reported changes in the relative abundance of subgroups within the *Acidobacteria*, an increase in *Bacteroidetes*, *Nitrospira*, *Alphaproteobacteria*, *Betaproteobacteria*, *Gammaproteobacteria*, and *Deltaproteobacteria* across the pH gradient. Variable responses to soil pH were also observed by Jones et al., (2009), demonstrating that although acidobacterial taxonomic diversity did not correlate significantly with pH, the relative abundance of operational taxonomic units (OTUs) associated with specific acidobacterial subgroups increased or decreased with decreasing soil pH. For example, acidobacterial subgroups 4, 6, 16, and 17 correlated positively with pH (with r values ranging from 0.74 to 0.91), whereas subgroups 1, 2, and 3 correlated negatively with soil pH (with r values ranging from -0.40 to -0.89).

Given the importance of microorganisms to soil fertility and biogeochemical cycling and the paucity of studies that have investigated soil pH and community composition, more work is required to identify soil bacterial taxa with abundances that correlate with pH, especially with data sets scaled to capture a large proportion of soil microbial community complexity. To do this, we used large 16S rRNA gene data sets generated by Illumina sequencing technology to examine composite soil samples from pH-gradient plots using multiple beta-diversity methods, including UniFrac-based principle coordinate analysis (PCoA) and non-negative matrix factorization (NMF; Jiang et al., 2011). Although UniFrac is a commonly used metric for measuring β-diversity (Lozupone et al., 2006), NMF has not been used for 16S rRNA gene data analysis prior to our study. NMF is useful in this context

as a data representation tool, whereby high-dimensionality data are converted to a few principle dimensions. After factorization, patterns of co-occurring OTUs can be described by a smaller number of taxonomic components. Each sample is represented by a collection of these component taxa, which help to display the relationship between taxa and the environment. Because the value of NMF for resolving patterns in metagenomic data sets was demonstrated only recently (Jiang et al., 2012; Jiang et al., 2011), we compared the results obtained from NMF to those from more common methods of 16S rRNA gene data reduction such as weighted and unweighted UniFrac. To verify the patterns observed in next-generation sequence data, we complemented this soil study with group-specific denaturing gradient gel electrophoresis (DGGE). Both sequencing and fingerprinting techniques demonstrated pH-dependent patterns within specific bacterial groups, both abundant and rare.

3.2 Materials and Methods

3.2.1 Soil sample collection

Craibstone Experimental Farm soil samples were collected from a defined agricultural soil pH gradient in Craibstone, Scotland (Scottish Agricultural Cottage; grid reference NJ872104; Podzol, sandy loam), where individual continuous plots have been maintained with seven discrete pH values (pH 4.5, 5.0, 5.5, 6.0, 6.5, 7.0, and 7.5) for over 50 years by the yearly addition of $Al_2(SO_4)_3$ or $Ca(OH)_2$ (lime). Soil plots were managed intensively to ensure high homogeneity. With all plots under the same 8-year crop rotation (winter wheat, potatoes, barley, root crop, oats, and grass for three years). Triplicate surface

soil samples (top 10 cm) were collected randomly from one soil gradient on two separate occasions (in late summer 2006 and September 11, 2007), and both were under potato crop each year. Replicate soil samples were sieved (3.25-mm) prior to storage at -80 °C. Soil chemistry and DNA extraction were conducted as described previously (Nicol et al., 2008), including measurements of soil pH with the $CaCl_2$ method.

Composite DNA samples were prepared by combining DNA extracts from each set of triplicate plot samples for each year (i.e. 7 pH subplots for each of 2006 and 2007, representing 14 composite DNA samples total). The composite DNA samples were used as template for Illumina sequencing with indexed primers, in addition to serving as template for DGGE fingerprinting as described above.

3.2.2 PCR-DGGE

Group-specific DGGE was conducted on all samples with primer sets and reaction conditions of Mülhling et al. (2008), with the exception of the primer sets used for *Acidobacteria* (Barns et al., 1999), *Verrucomicrobia* (Stevenson et al., 2004), and *Actinobacteria* (Stach et al., 2003), where the PCR was conducted according to methods in the corresponding publications. Gels consisted of 8% acrylamide and bis-acrylamide (37.5 : 1), with a denaturing gradient from 40% to 60% (100% denaturant contains 7 M urea and 40% formamide). Equal amounts of PCR product, measured to the ng, were loaded into each well, and gels were run at 85 V for 14 h. An in-house ladder was run on each gel, helping with profile normalization. After post-staining with SYBR Green I, gels were imaged on either a Typhoon 9400 Variable Mode Imager (GE Healthcare, Waukesha, WI) or a Pharos

FX Imager (Bio-Rad, Hercules, CA). Gelcompar II (Applied Maths, Austin, TX) was used to normalize gels and generate dendrograms based on Pearson's correlations of densitometric curves. The data was clustered using unweighted pair group method with arithmetic mean for the group-specific DGGE gel fingerprint dendrograms.

3.2.3 qPCR

Bacterial 16S rRNA gene abundance was assessed for each composite soil sample using the primer set 341f and 518r (Muyzer et al., 1993).Quantitative PCR was run on a CFX96 (Bio-Rad) PCR machine. Each PCR mixture contained 6 µL of iQ SYBR Green Supermix (Bio-Rad), 0.4 µM of both forward and reverse primers, 5 µg of bovine serum albumin, and ~0.5 ng of DNA template. Standard DNA was generated using extracted and quantified soil DNA and was amplified using the primers 27f and 1492r (Lane 1991). A serial dilution of standard was added as template to the qPCR to generate a standard curve. PCR conditions used were 95°C for 10 min, followed by 95°C for 30 s, 55°C for 30 s, and an elongation step at 72°C for 30s, which was repeated 40 times. Each elongation was concluded with a fluorescent plate read. The coefficient of determination of the standard curve was 0.99, and the efficiency was 85%.

3.2.4 Illumina library generation and sequencing

Illumina-based PCR amplification and cycle conditions were the same as those detailed in section 2.2.2 of this thesis. Briefly, 10 ng of DNA from each composite sample was added to triplicate PCR amplifications, 20 cycles for each sample, and the products of these replicate reactions were size-selected on an agarose gel, purified, and pooled to generate composite amplicon templates. Pooled amplicon templates were analyzed for

concentration and size by agarose gel electrophoresis, absorbance (NanoDrop; Thermo Scientific), and microfluidics (Bioanalyzer; Agilent). Paired-end sequencing (2 x 125 bases; 6-base index read) was performed on a Genome Analyzer IIx (Illumina) with version 4.0 sequencing reagents.

3.2.5 Bioinformatic analysis

The CASAVA pipeline (version 1.6) was used for base calling and error estimation of sequence reads. Following this initial quality-control step, primer-free 150 base paired-end reads were assembled as in Chapter 2 of this thesis, using a prototype version of PANDASEQ (Masella et al., 2012). Briefly, sequences with ambiguous bases or mismatches in the overlap region (12-base minimum overlap) were removed, in addition to removing sequence regions corresponding to PCR primers. Following assembly, sequences containing fewer than 75 bases were excluded. Using the QIIME software package (Caporaso et al., 2010b), managed by AXIOME (Lynch et al., 2013), taxonomy was assigned to each sequence using the naeive Bayesian classifier of the Ribosomal Database Project (RDP-II) (Wang et al., 2007) with an assignment cutoff of 0.5, which had been shown previously to be appropriate for short sequence reads (Claesson et al., 2009). Following this, sequences were aligned (PYNAST;(Caporaso et al., 2009)), and a phylogeny was constructed. Of the 2,033,920 sequences that were reduced to 23 088 clusters, 1080 OTUs (equaling 7828 or 0.38% of sequences) were not aligned. A PCoA ordination was plotted using both weighted and unweighted UniFrac distances (Lozupone et al., 2006; Lozupone et al., 2010). The NMF analysis was conducted according to the methods of Jiang and coworkers (Jiang et al., 2012), using rarefied OTU profiles (clustered at 97% identity). NMF is sometimes used for

clustering, but here we use it for dimensional reduction. The NMF factorization of OTU profile can be thought of as an empirical attempt to describe observed OTU patterns according to a small number of taxonomic "components". The observed OTU distribution for each sample is represented by a weighted sum of component abundance distributions. Similarities between OTU abundance distributions and NMF component profiles were calculated as described previously (Jiang et al., 2012b). All NMF analyses were conducted on a desktop computer, and the R code for this analysis is available here: http://lalashan.mcmaster.ca/theobio/soil_metagenomics/ index.php/Ph_nmf.

Based on the concordance analysis, we chose to examine component taxa associated with rank 3 and rank 5 decompositions. Taxonomy for representative component taxa was visualized using SSUnique (Lynch et al., 2012). Briefly, nodes corresponding to representative taxa were connected edgewise to a central (square) node defining RDP-assigned taxonomy (confidence of > 0.5), visualizing taxonomic consistency within the data, that is, unconnected OTU nodes were not assigned to established taxonomies at the confidence threshold. Sequence data were deposited in the Sequence Read Archive (SRA; NCBI) with the accession number SRP007517.

3.3 Results

A total of 14 composite soil samples (i.e. 7 samples from each of 2 years) from the Craibstone Experimental Farm plots were characterized by measuring soil chemistry, and measuring the microbial community via gel fingerprinting, and sequencing of the bacterial

16S rRNA gene V3 region. Soil chemistry demonstrated that the defined pH values for each plot were similar to the measured pH values for those same plots in both 2006 and 2007 (Table 3.1). Although no observable pH-dependent or year-dependent trends were visible for organic carbon and moisture content, we observed yearly differences in ammonia and nitrate concentrations in the subplots even though N : P : K fertilizer was applied consistently each year to the potato plots at a rate of 100:150:120 kg ha^{-1}. Overall, bacterial 16S rRNA gene relative abundance increased with increasing pH ($r = 0.817$; Table 3.1); this trend was observed for both 2006 and 2007.

Sequence data were used in conjunction with DGGE to characterize bacterial composition and diversity present in composite soil samples ranging from pH 4.5 to 7.5. DGGE also provided initial justification for sequencing composites from replicate soil samples because replicate soil DNA extracts generated nearly identical fingerprints for each plot and time point (Fig. 3.1). Final assembly of 7,536,750 paired-end Illumina sequences contributed 146,087–888,148 assembled 150-base contigs per composite sample. Alpha diversity was measured for each sample from a rarefied data set of 146,087 sequences from each sample. As expected, Chao1, Shannon diversity, and the number of observed species were highest in samples of high pH (r values of 0.686, 0.764 and 0.750 respectively; Table 3.1). Good's coverage (Good, 1953) ranged from 0.981 to 0.992 and, in general, decreased with increasing pH (Table 3.1), reflecting higher diversity with increasing pH.

Table 3.1 Composite soil sample chemistry and bacterial community data for two sample years (2006 and 2007), with sequence analysis based on data rarefied to 146,087 sequences per sample.

Set pH	Measured pH	Year	%C	%N	Ammonium ($\mu g\ g^{-1}$ soil)	Nitrate ($\mu g\ g^{-1}$ soil)	Bacterial qPCR ($x10^9$copies g^{-1} soil)	Bacterial qPCR Standard Deviation ($x10^8$)	Sequences per sample	Shannon index	Chao1	Observed OTUs	Good's coverage
4.5	4.4	2007	6.02	0.58	5.80	0.79	1.73	1.08	473813	8.5	5462	3964	0.991
5.0	4.9	2007	5.60	0.56	6.36	4.83	2.27	1.77	730697	9.1	8081	5697	0.986
5.5	5.5	2007	6.02	0.61	7.68	0.99	5.18	1.41	757394	9.7	9926	7058	0.983
6.0	6.0	2007	5.86	0.59	7.46	0.95	4.38	4.80	889584	9.7	10725	7269	0.981
6.5	6.6	2007	5.87	0.55	8.57	1.42	7.93	0.25	146087	10.0	8208	6877	0.988
7.0	7.0	2007	5.44	0.52	8.52	1.10	6.70	2.56	420812	9.9	9756	7101	0.983
7.5	7.3	2007	5.57	0.60	8.17	2.17	7.53	0.25	621783	10.1	10863	7872	0.981
4.5	4.9	2006	7.02	0.38	1.41	15.07	1.24	2.59	477403	8.3	5291	3860	0.991
5.0	5.3	2006	6.60	0.37	1.47	14.94	1.05	1.43	172354	8.6	5208	4284	0.992
5.5	5.9	2006	7.42	0.40	1.65	10.19	2.91	0.82	503986	9.0	7996	5654	0.986
6.0	6.4	2006	6.57	0.31	1.42	9.48	4.43	2.75	781626	9.2	9081	6173	0.984
6.5	6.9	2006	6.42	0.29	1.36	9.77	5.61	1.07	551830	9.5	9078	6463	0.984
7.0	7.3	2006	7.97	0.36	1.46	10.46	5.17	3.90	407709	9.5	8733	6405	0.985
7.5	7.5	2006	7.14	0.34	1.71	10.99	5.50	0.77	627541	9.6	9338	6660	0.984

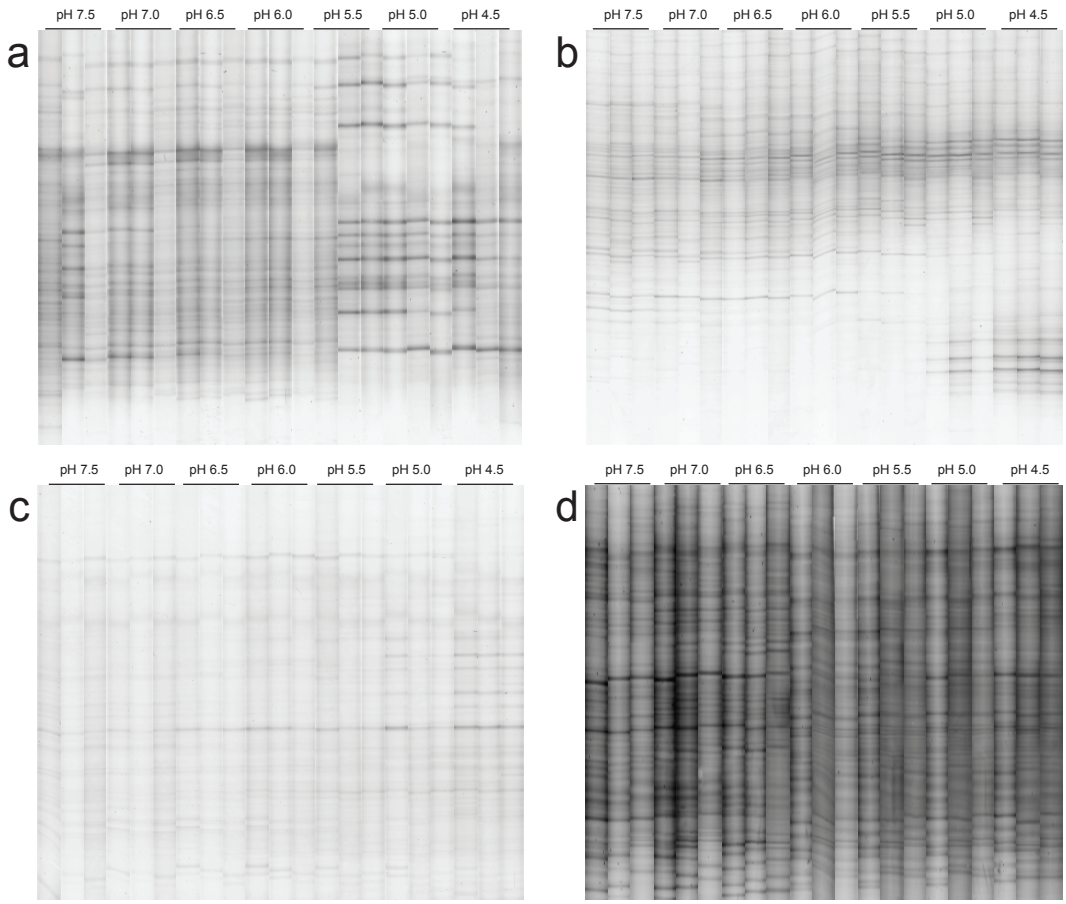

Figure 3.1 Single-year 16S rRNA gene DGGE profiles of triplicate soil samples across the pH gradients using group-specific and general bacterial primers for (a) *Acidobacteria*, (b) *Verrucomicrobia*, (c) *Alphaproteobacteria*, and (d) Bacteria.

3.3.1 Community composition

Bacterial and group-specific DGGE was used to assess bacterial composition associated with composite soil samples from each defined pH plot (Fig. 3.2). General bacterial DGGE patterns revealed complex communities, with only a subtle pH-dependent shift in band diversity and composition (Fig. 3.2a). On the other hand, DGGE fingerprints for

several individual bacterial phyla and subdivisions demonstrated pronounced shifts in community fingerprints across the pH gradient (Fig. 3.2b–h). Based on Pearson correlations of densitometric curves, fingerprints from all targeted groups clustered according to pH, with low-pH soil fingerprints clustering separately from high pH soil fingerprints. *Acidobacteria*, *Verrucomicrobia*, and *Gammaproteobacteria* exhibited the most pronounced changes, with unique DGGE patterns associated with composite soil samples from pH 4.5 and 5.0 plots. Fingerprints for *Firmicutes*, *Actinobacteria*, *Alphaproteobacteria*, and *Betaproteobacteria* also showed pattern changes across the gradient, but to a lesser extent. All of the observed trends were consistent for both 2007 (Fig. 3.2) and 2006 (Fig. 3.3) soil sample analyses.

In general, taxonomic affiliations of the Illumina sequence data corroborated these initial DGGE observations (Fig. 3.2, Fig. 3.3). Importantly, the shift in acidobacterial community composition was greatest for both sequence data and Pearson correlations of densitometric curves of the corresponding DGGE fingerprints. Although plotted, phylogenetic representation of sequence data at the phylum level revealed no clear associations with pH (e.g. acidobacterial groups were associated with both low and high pH; Fig. 1), the relative abundance of *Acidobacteria* subgroups 1, 2, and 3 increased at low pH, concomitant with a proportional decrease in the representation of subgroups 4, 6, 14, and 16. Trends were consistent for plotted ordinal taxonomic affiliations of the Illumina-generated sequences for 2007 (Fig. 3.2) and 2006 (Fig. 3.3).

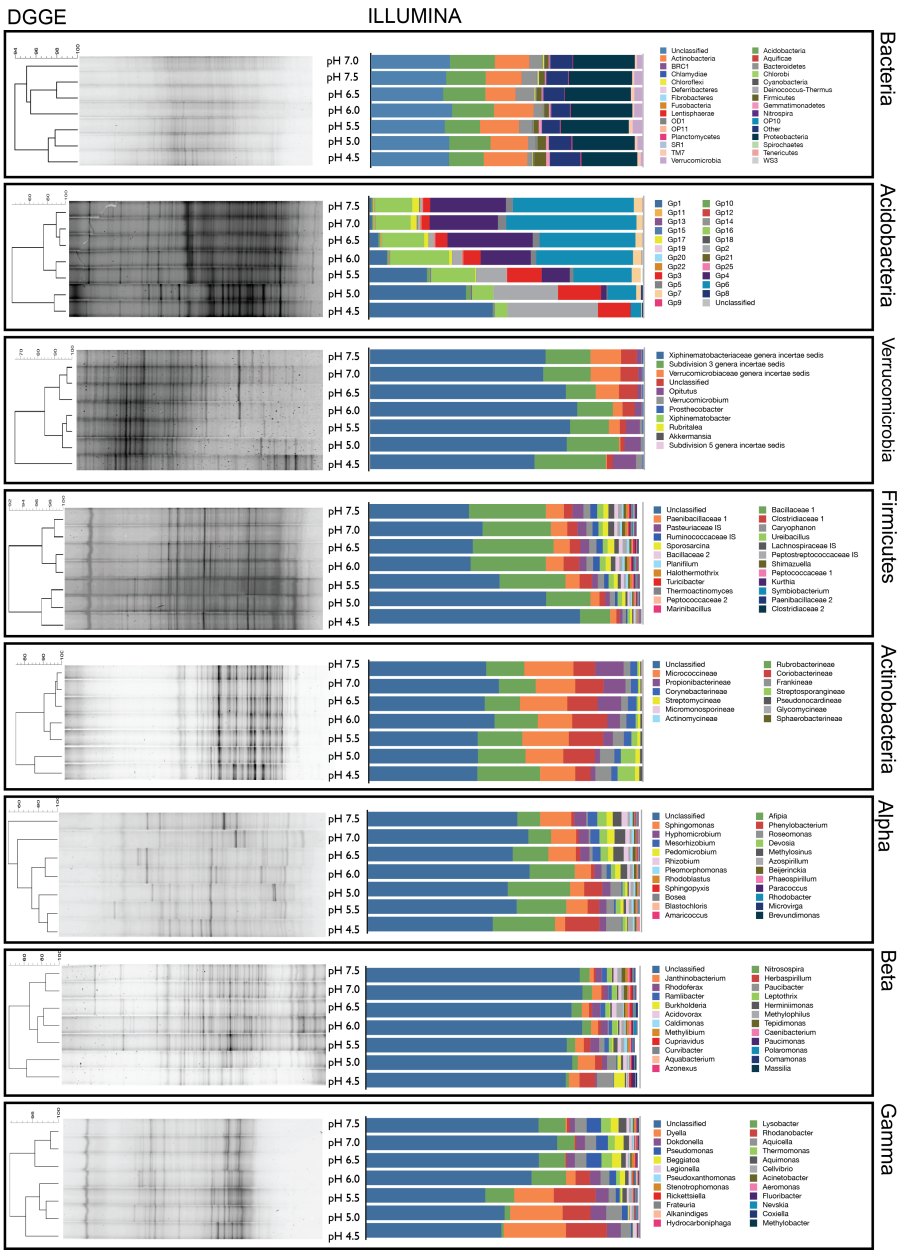

Figure 3.2 DGGE group-specific profiles of composite soil samples (left) from 2007 with the corresponding taxonomic proportions taken from the Illumina sequence library. Plotted proportional abundance only took into account the 25 most abundant taxa (if applicable). (a) *Acidobacteria*, (b) *Verrucomicrobia*, (c) *Firmicutes*, (d) *Actinobacteria*, (e) *Alphaproteobacteria*, (f) *Betaproteobacteria*, and (g) *Gammaproteobacteria*.

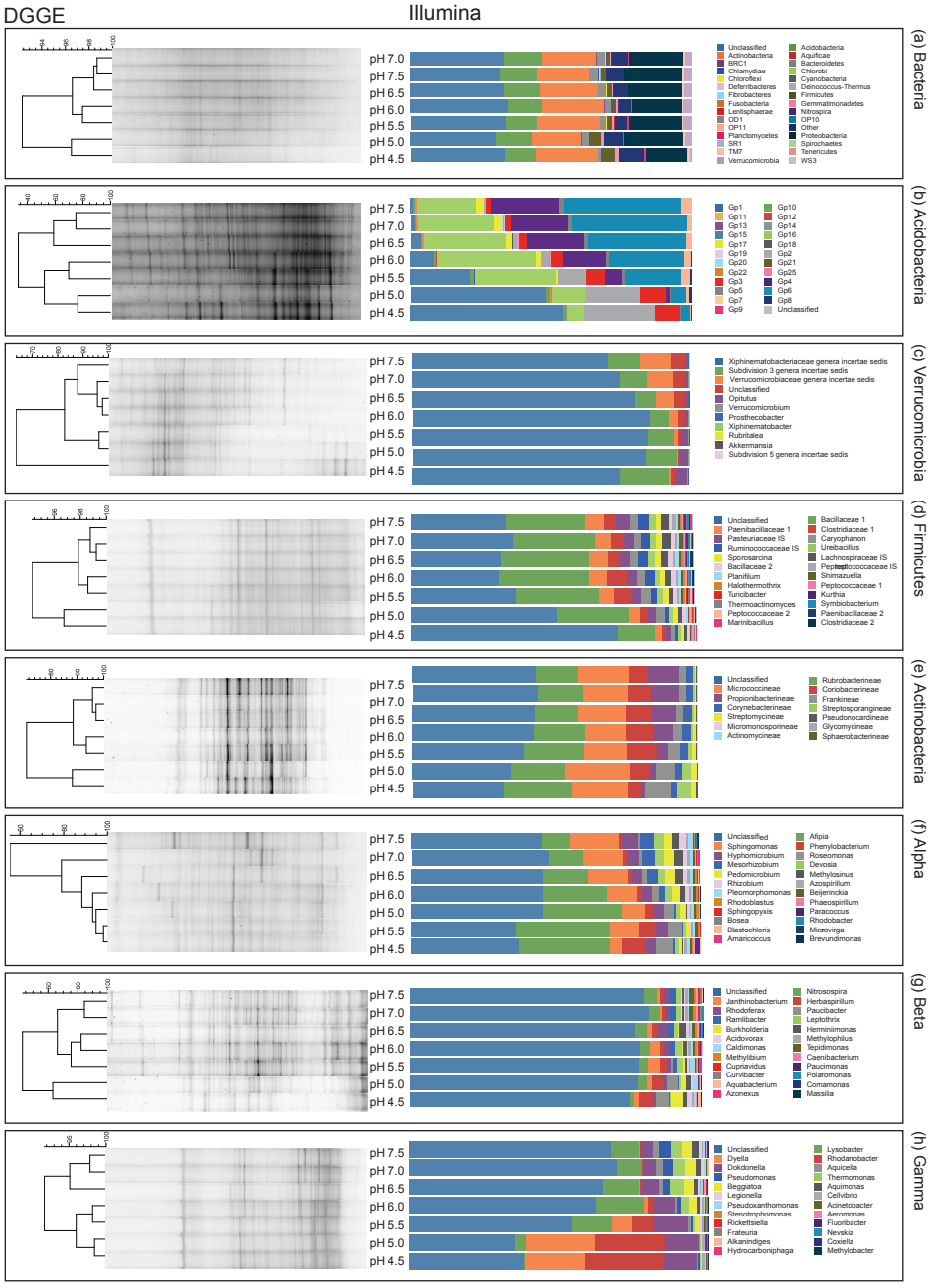

Figure 3.3 DGGE group-specific profiles of soil samples (left) from 2006 with the corresponding taxonomic proportions taken from the Illumina sequence library. Plotted proportional abundance only took into account the top 25 most abundant taxa (if applicable). (a) *Acidobacteria*, (b) *Verrucomicrobia*, (c) *Firmicutes*, (d) *Actinobacteria*, (e) *Alphaproteobacteria*, (f) *Betaproteobacteria*, and (g) *Gammaproteobacteria*.

3.3.2 Beta diversity

Both PCoA and NMF were used to characterize the response of soil bacterial communities to pH, in addition to secondary soil characteristics that differed over the 2-year period and between plots (e.g. ammonia and nitrate; Table 3.1). The UniFrac distance metric was used to measure between-sample phylogenetic distances for preparing PCoA ordination plots. The unweighted and weighted UniFrac PCoA plots for all 2006 and 2007 composite samples clustered by pH (Fig. 3.4a and b, respectively), with the unweighted plot exhibiting tighter clustering for samples of similar pH (i.e. samples are closer together on axis 2). Unweighted UniFrac-based PCoA plots were prepared with taxonomic biplot overlays for phylum, class, and genus levels (Fig. 3.5). The position of each taxon 'bubble' indicates that particular taxon's relative importance in contributing to the sample's position in the plot. Many of the same trends observed in the DGGE and taxonomic affiliations plots (Fig. 3.2) were confirmed by the biplots. For example, *Acidobacteria* subgroups 1, 2, and 3 are proximal to the low-pH samples; subgroups 4, 6, and 16 cluster with the high-pH samples (Fig. 3.5c).

We complemented PCoA plots of UniFrac distances with NMF analysis, which is a multivariate method for identifying 16S rRNA gene b-diversity patterns and retrieving co-occurring positively interacting components of complex datasets.

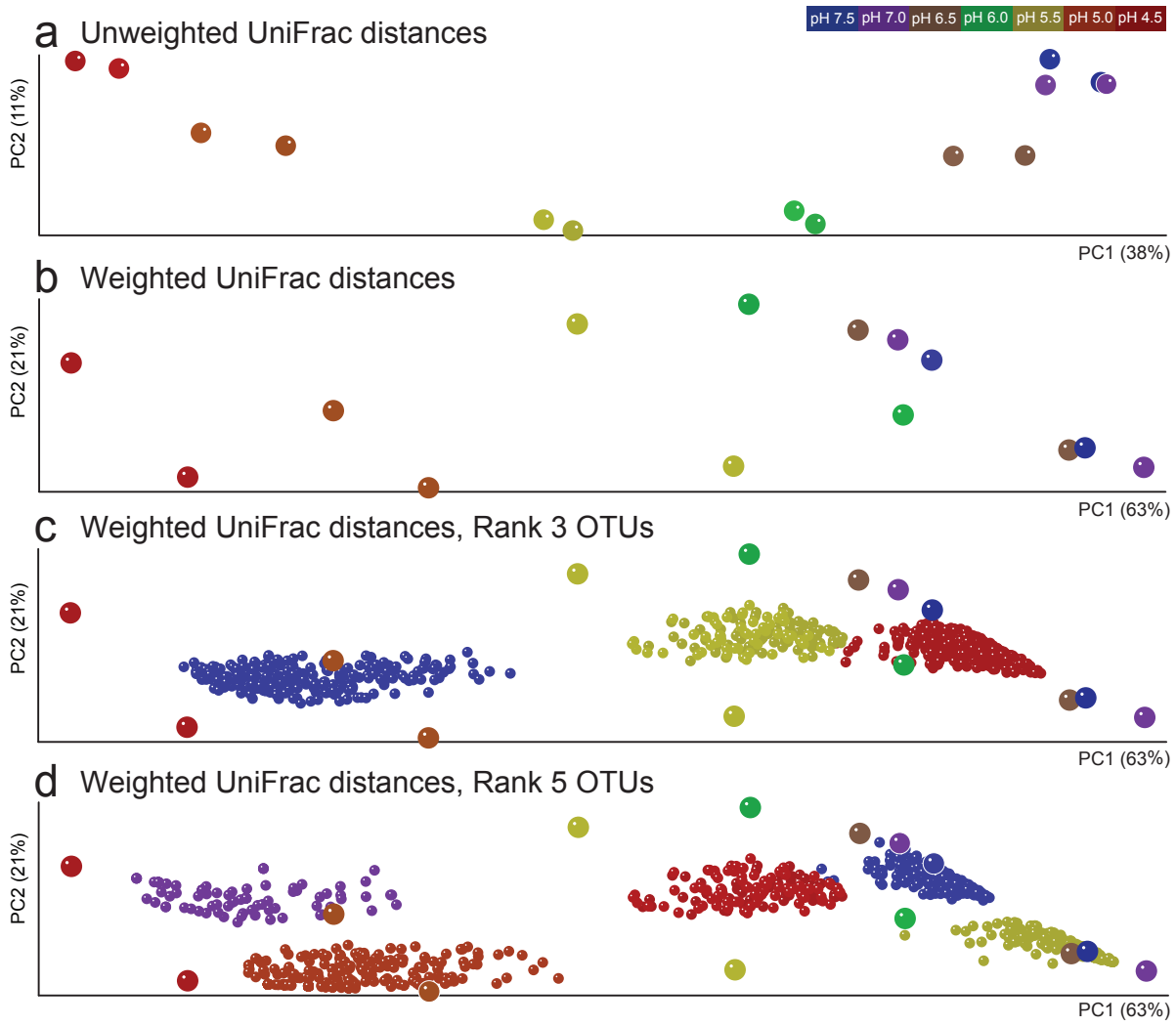

Figure 3.4 Clustering of sequence data for 2006 and 2007 composite soil samples from the Craibstone Experimental Farm. Principle coordinate analysis (PCoA) is based on unweighted UniFrac distances (as shown horizontally mirrored to Fig. 3.5; a), weighted UniFrac distances (b), and weighted UniFrac distances with a superimposed plot of NMF rank 3 representative OTUs (c). The spheres correspond to representative taxa for high (red), medium (yellow) and low (blue) pH. A PCoA plot based on weighted UniFrac distances is also shown for NMF rank 5 representative OTUs (d). These spheres represent representative taxa for NMF components 1 (blue), 2 (yellow), 3 (red), 4 (purple), and 5 (orange).

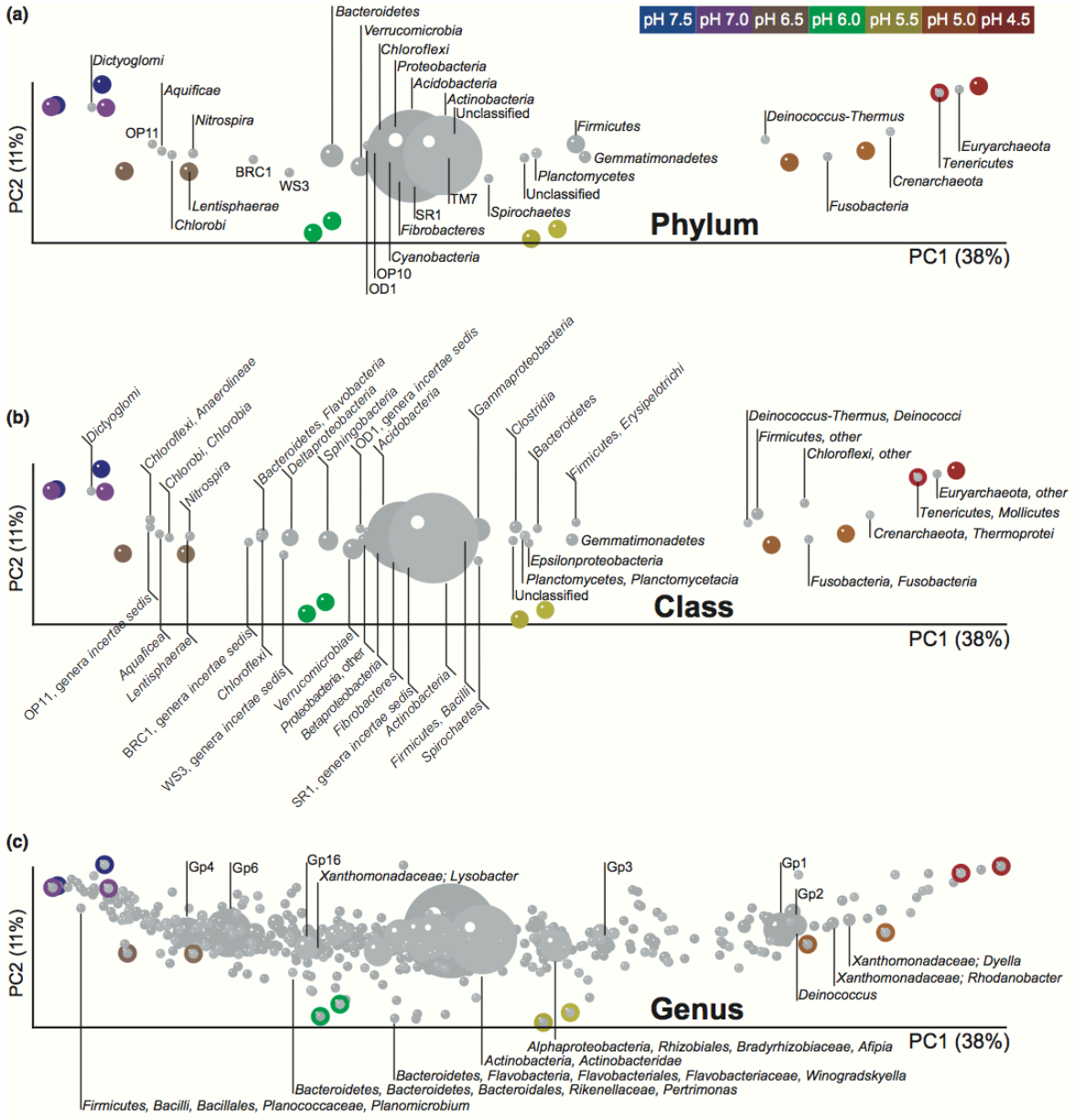

Figure 3.5 Clustering of sequence data for 2006 and 2007 composite soil samples from the Craibstone Experimental Farm. PCoA is based on unweighted UniFrac distances; biplot overlays demonstrate taxa that contributed to sample differentiation at the phylum (a), class (b), and genus levels (c), and their relative size corresponds to number of summarized taxa belonging to that group. Percent of data variability explained by each axis is indicated.

A rarefied OTU table (i.e. sequences were selected pseudorandomly from each sample down to the size of the smallest sample library) was used as input for the factorization process, with 146,087 sequences per sample. Based on the NMF concordance method, ranks 2, 3, and 5 decompositions showed strong local peaks (Fig. 3.6a). Ranks 3 and 5 were selected for further analysis. Correlations of NMF-based sample similarity matrices and chemical parameters were also plotted, indicating that component OTUs representing the rank-3 decomposition were strongly associated with pH (Fig. 3.6b, Fig. 3.7). Strong clustering based on soil pH was also evident when representative OTUs associated with the rank-3 NMF decomposition were superimposed on UniFrac-based PCoA plots (Fig. 3.4c).

The OTU clusters associated with a rank-5 NMF decomposition were correlated with high pH (clusters 2 and 3), low pH (clusters 4 and 5), medium pH (cluster 1), year, and also with nitrogen concentration (Fig. 3.6c). Note that nitrogen concentrations differed between 2006 and 2007 (Table 3.1), which may explain the observed association of both nitrogen and sampling year with bacterial community composition. The sample similarity matrices, discussed previously (Figs 3.6b and c), were also used to create a heat map containing representative component taxa (non-negative correlating taxa) related to the samples (Fig. 3.6d and e). Representative OTUs from the rank-5 NMF decomposition were visualized as a biplot overlay on the UniFrac-based PCoA plots (Fig. 3.8d), with the rank-5 taxa showing vertical spread, likely due to the influence of the sample year.

At the family level, for the rank-3 cluster taxonomic affiliations (Fig. 3.8), unclassified sequences comprised a higher proportion of the low-pH cluster (41.5%)

compared with the intermediate (24.8%) and high-pH clusters (36.0%). Other phyla, such as *Verrucomicrobia*, increased in relative abundance in the medium-pH cluster. *Acidobacteria* increased in the high-pH cluster, and *Actinobacteria* were a higher proportion of the low-pH cluster (Fig. 3.8a). For many groups, shifts in taxonomic composition were more pronounced above order (Fig. 3.8), and consistencies were observed between taxa present in NMF and those in the PCoA biplots. For example, *Acidobacteria* Gp 4, 6, and 16 were associated with medium to high-pH clusters, and Gp 1, 2, and 3 were associated with low-pH clusters, shown by both NMF and PCoA biplots (Fig. 3.5c, Fig. 3.8).

When other taxonomic groups were considered, several taxa were only observed within a particular NMF cluster. This includes the genera *Burkholderia* and *Paucibacter* (*Betaproteobacteria*), which were found within the NMF rank-3 low-pH cluster only, and *Leptothrix* only being found in the medium-pH cluster. Sequences classified within the genera *Nitrosospira*, *Denitratisoma*, *Paucimonas*, *Herbaspirillum*, *Tepidimonas*, and *Polaromonas* (*Betaproteobacteria*) were associated with the high-pH cluster. Within the Alphaproteobacteria, the genus Phenylobacterium corresponded to the low-pH cluster, while the genera *Devosia*, *Roseomonas*, *Labrys*, *Methylosinus*, *Fulvimarina*, *Filomicrobium*, *Rhodobacter*, *Hyphomicrobium*, *Bartonella*, and *Mesorhizobium* were associated solely within the high-pH cluster (Table 3.7, 3.8). Within *Gammaproteobacteria*, the genera *Dyella* and *Rhodanobacter* (both classified to the *Xanthomonadaceae* family) were only found within the low-pH NMF rank-3 decomposition and were located toward the low-pH samples within PCoA biplots (Fig. 3.5c and Table 3.2). Conversely, within the same family, *Lysobacter* was observed within the high and medium-pH clusters only and was proximal to

the high-pH sample on the PCoA biplot (Fig. 3.5c and Table 3.2).

To visualize both the relative abundance and taxonomic affiliations of the pH-dependent representative taxa identified by NMF, a network diagram was prepared with nodes representing rank-3 OTUs and edges representing familial identities to the closest representative sequence in the RDP-II database (Fig. 3.9). In addition to summarizing the taxonomic affiliations of the NMF rank-3 representative taxa, the size of the nodes represents the abundance of each OTU in each rarefied sample data set. The number of OTUs in each rank-3 NMF decomposition varied, with the medium-pH cluster containing the fewest OTUs and sequences (137 and 35,375, respectively). The high-pH cluster contained the most OTUs and sequences (551 and 279,176 respectively), and the low-pH cluster contained 144 OTUs represented by 238,559 sequences. For the low-pH NMF component taxa, a total of 129 OTU nodes (representing 139,674 sequences) were connected (classified to the familial), yet 144 OTUs (representing 98,885 sequences) were represented by unconnected nodes. The medium-pH cluster contained 86 connected OTUs (26,615 sequences) and 51 unconnected OTUs (8760 sequences). The high-pH cluster was the largest, with 305 connected OTUs (178 207 sequences) and 246 unconnected OTUs (100,969 sequences). The low-pH cluster contained the highest proportion of unconnected OTU nodes at 52%, compared with the medium and high clusters with 37% and 45% unconnected nodes, respectively. For the low-pH cluster, this translates to 41% of the total sequence reads that were unconnected, because unclassified OTUs were predominantly associated with low-abundance taxa.

.

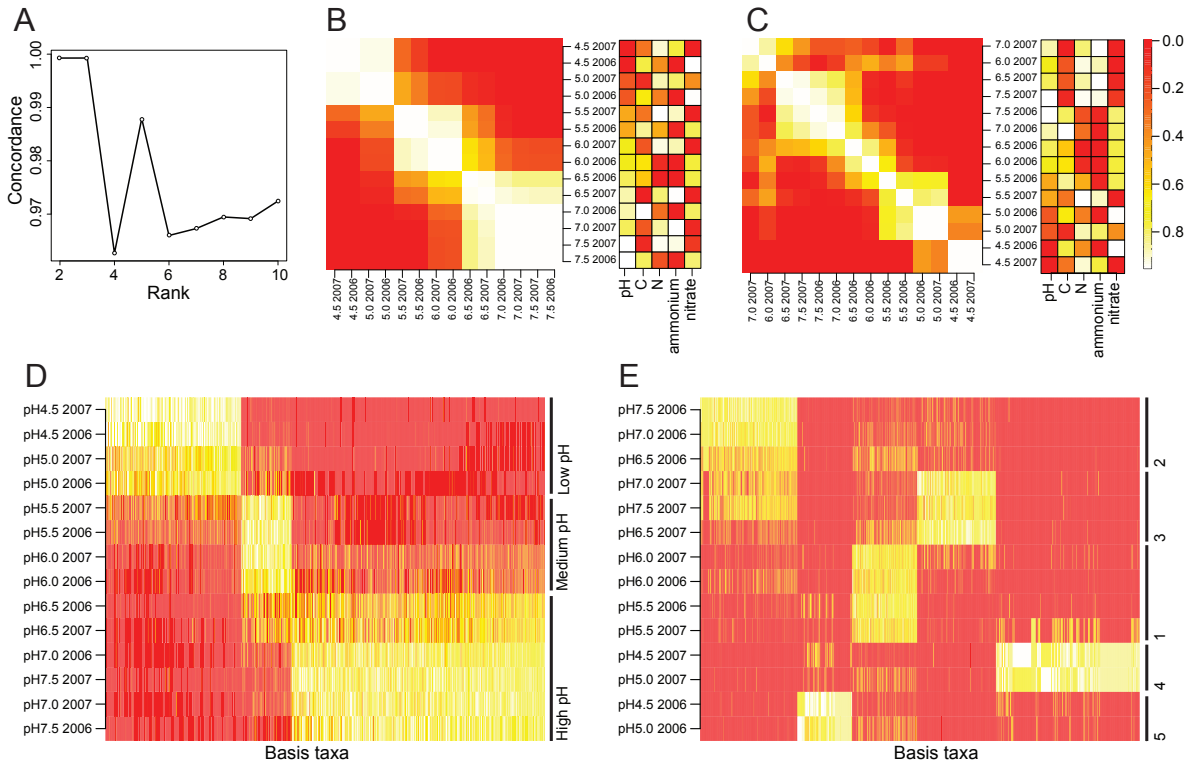

Figure 3.6 NMF heat maps for rank 3 and 5 decompositions. Concordance model (a) demonstrates rank 3 and 5 as stable ranks for NMF. Sample similarity matrices are shown with correlations to chemical parameters for rank 3 (b) and rank 5 (c). Abundance distributions of NMF representative taxa are shown for sample plots for rank 3 (d) and rank 5 (e).

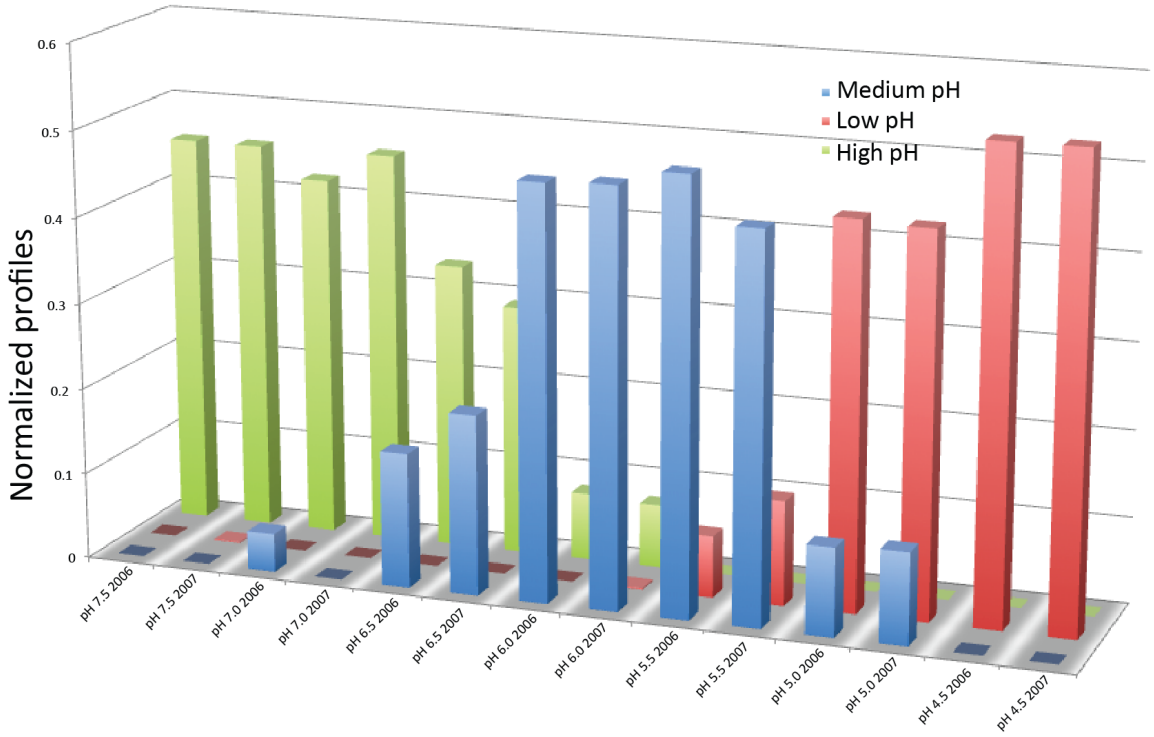

Figure 3.7 Normalized relative abundance of three NMF components at rank 3 for each soil sample, demonstrating the strong affiliation of rank 3 components and pH. The three components are coded by three different colors (red, blue, and green), normalized by dividing by the sum of each component's total.

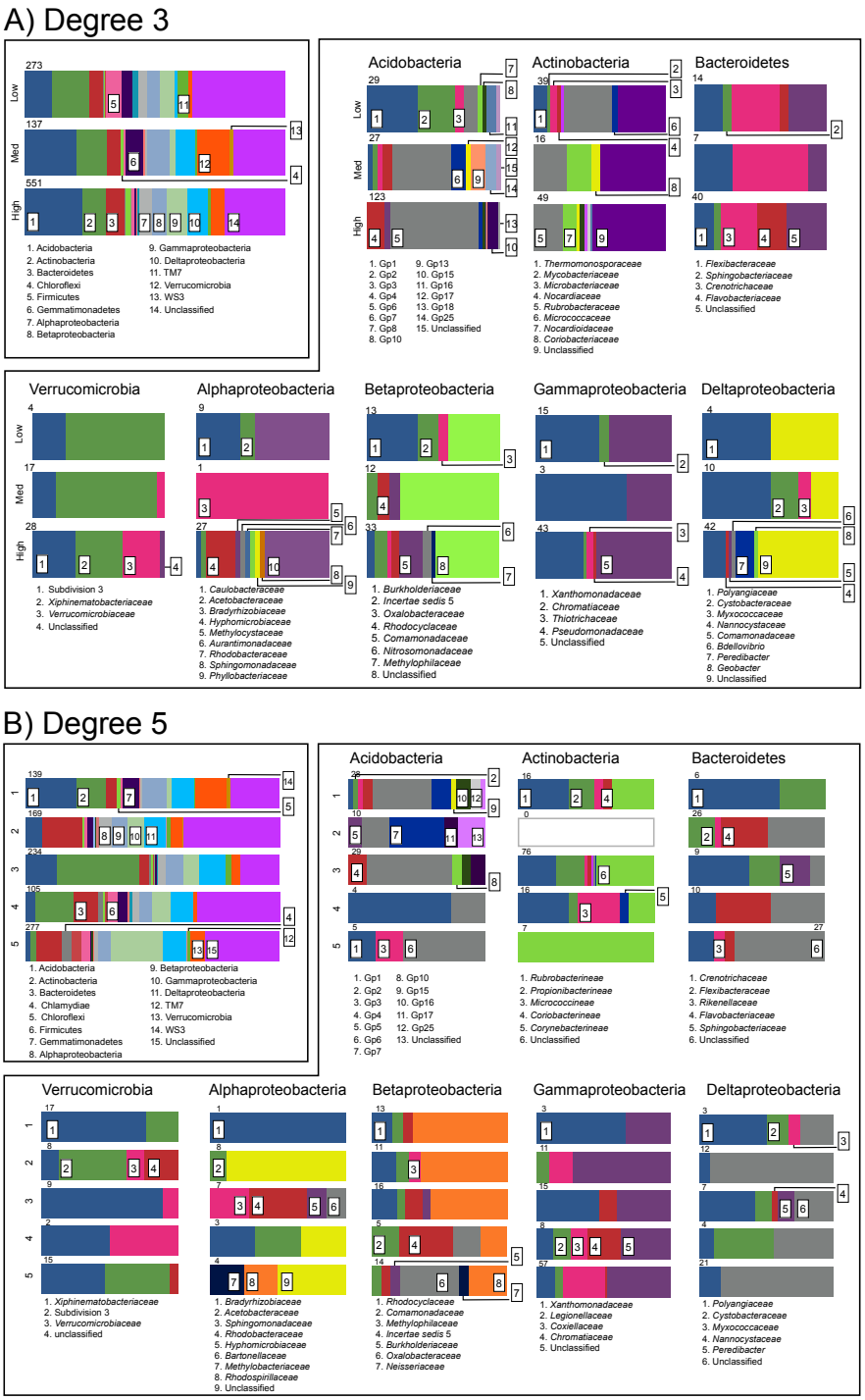

Figure 3.8 Taxonomic affiliations of rank 3 representative OTUs (a) and rank 5 representative OTUs (b). Numbers on the top left of each bar represent the number of OTUs found within the NMF-based components that were used to make the graph

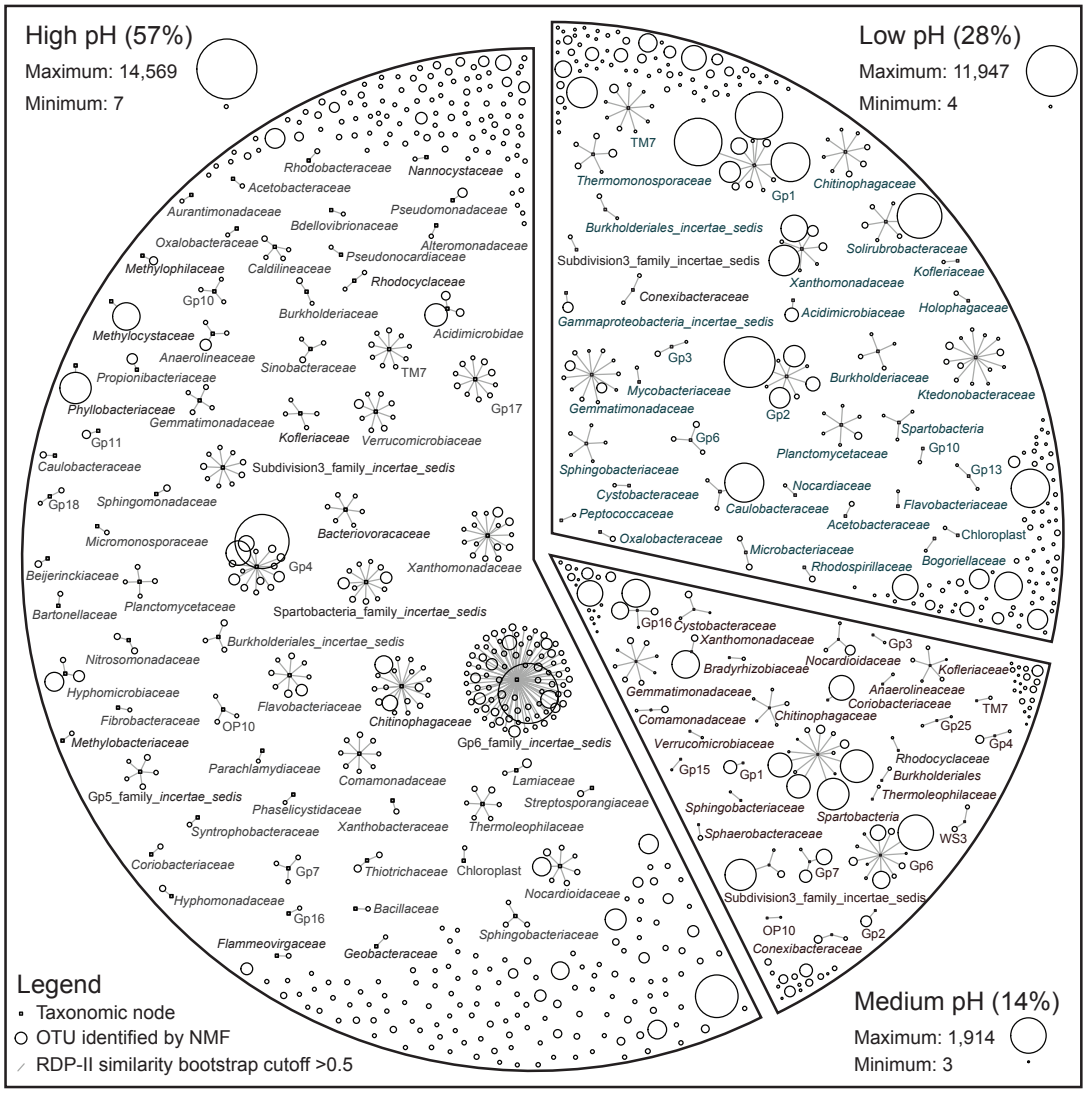

Figure 3.9 Connections of representative NMF component OTUs and characterized taxa within the RDP-II, represented as a network visualization of familial connections for high pH, medium pH, and low pH. OTU sequences were connected to taxonomic nodes by edges if there was a corresponding RDP-II classification with a bootstrap cutoff of ≥ 0.5. The diameter of OTU nodes is a linear function of the sequence abundance of that OTU. Representative OTU components for each rank were processed independently, and a composite figure was manually generated. Gp, Acidobacterial subgroup.

3.4 Discussion

Our research adds to recent literature demonstrating the important influence of pH on soil diversity and bacterial community composition. In addition to complementing previous research, our study adds novelty in four ways: (1) this is one of only two studies of pH influences on bacterial diversity associated with experimental agricultural plots; (2) we generated sequence data sets that are orders of magnitude larger than all previous studies combined; (3) we introduce novel multivariate and taxonomic analyses (i.e. NMF, SSUnique) to expand our knowledge of pH-based effects on soil microbial communities; and (4) we show that inexpensive fingerprinting methods show similar results to Illumina sequencing at discerning differences in communities at a coarse level.

The Craibstone Experimental Farm pH plots were selected for this study because individual soil plots possess similar physical and chemical characteristics, which enabled the effects of pH to be considered independently from other soil physicochemical properties (Table 1). Building on past Craibstone soil plot observations that soil pH influences the distribution of ammonia-oxidizing archaea (Lehtovirta et al., 2009; Nicol et al., 2008), we hypothesized that soil pH would associate with bacterial diversity and community composition and that large sequence data sets would reveal both abundant and rare taxa that correlate with pH. Prior to the advent of next generation sequencing, Fierer and Jackson (2006) used terminal restriction length polymorphism of North American soil samples to identify the link between soil pH and bacterial diversity, with a diversity maximum at neutral pH. Other factors such as annual temperature, potential evapotranspiration, and latitude were

found to be poor overall predictors of microbial diversity. Although several subsequent studies have extended these initial findings (e.g. Chu et al., 2010; Jones et al., 2009; Lauber et al., 2009), only one other study of soil pH has investigated the effect of soil pH using experimental plots (Rousk et al., 2010). Because soil pH is the primary factor influencing soil bacterial diversity and composition, multiple studies from varied soil sample collections and study sites are important for reinforcing findings associated with individual studies. Here, we used the only published Illumina 16S rRNA gene sequencing approach coupled with paired-end read assembly (Chapter 2) to generate high numbers of reads per sample (a range of 146 087–889 584 sequences per sample), two to three orders of magnitude more data than all past soil pH studies combined. Rousk et al., (2010) and Lauber et al., (2009) used either 600 rarefied sequences or an average of 1501 sequences per sample, respectively. This current increase in sequencing depth (Good's coverage > 0.98 for all samples; Table 3.1) was useful for identifying even low-abundance OTUs associated with pH.

NMF was used as a representation method for portraying high-dimensional data as a small number of taxonomic components. The observed OTU distribution of a sample is represented approximately by a weighted sum of component abundance distributions. Like principal component analysis (PCA), NMF decomposes an input matrix into components, with the goal of making a low-dimensional approximation. Unlike PCA, NMF is an approximate decomposition, but it has the advantage that both the components and their contributions are non-negative (positive, or zero). Also unlike PCA, the NMF decomposition is dependent on the number of components used (the rank). Mathematically, if we have p OTUs and s samples, then the size of the profile matrix X is p 9 s. NMF finds matrices W

and H (with dimension p 9 k and k 9 s, respectively, where k is the rank of our factorization), such that WH = X. We search for the approximations that minimize the Kullback–Leibler (KL) divergence between X and WH (Jiang et al., 2012b).

Because NMF is a mathematical analysis method for representing high-dimensional data as positive linear combinations of positive components (Jiang et al., 2012; Jiang et al., 2011), NMF was ideally suited for this study due to its ability to resolve patterns in large 16S rRNA gene sequence data sets, informed by a concordance model. The majority of the OTUs we identified within NMF clusters were low-abundance sequences, occurring c. 10 times, illustrated by small circles that were frequently unconnected to the RDP reference taxonomic backbone using SSUnique analysis (Fig. 3.9; (Lynch et al., 2012)). NMF clusters containing fewer than 30 sequences ranged from 9% to 15% of all OTUs identified within each cluster (Fig. 3.9). Importantly, these represent OTUs that would not have been detected by smaller data sets used commonly for beta-diversity analyses in microbial ecology.

The assembled paired-end Illumina data confirmed previous observations that bacterial diversity is lowest in acidic soil samples and soil diversity reached a maximum at pH 7.5 (Table 3.1). These pH-specific results were consistent for both 2006 and 2007. Despite this interannual consistency, the weighted UniFrac PCoA Craibstone plots (Fig. 3.4) revealed more separation between duplicate pH soil samples based on year, compared with unweighted UniFrac plots (Fig. 3.5). This indicates that although pH had a strong effect on the presence/absence of OTUs (unweighted UniFrac) for both 2006 and 2007 samples, the relative abundance of those OTUs varied somewhat from year to year (weighted UniFrac) in

addition to pH-based sample separation. This is likely due to approximately the same species either being present or absent in the corresponding pH plot regardless of year, and with the numbers of each being more variable on a yearly basis (explaining the much larger x-axis variability of 63%). This is evidence that despite best efforts to keep all other parameters consistent save pH, there were other temporal, chemical, and/or physical influences involved in governing microbial community composition.

Acidobacteria are found in many environments and possess diverse metabolic functions, with certain subgroups being related to specific soil conditions such as temperature, carbon content, and pH (Rawat et al., 2012). Despite their numerical importance, little is known regarding this group's distribution, function, and overall contribution to soil ecosystems. The NMF analysis (Fig. 3.6, 3.7, 3.8) demonstrated that phyla such as the *Acidobacteria* are represented by taxonomic groups that are associated with high, medium, or low pH, suggesting that specific acidobacterial subgroups are adapted to distinct pH conditions. For example, OTUs affiliated with *Acidobacteria* subgroups 1, 2, 3 were more abundant at a lower pH, and those affiliated with subgroups 4, 16, and 6 were more abundant at neutral pH (Fig. 3.8). These trends in relative abundance were very similar to those found in previous studies (Jones et al., 2009; Rousk et al., 2010), with subgroup 1 associated primarily with acidic soil samples (Jones et al., 2009; Rawat et al., 2012). Additional acidobacterial subgroups identified within the NMF rank 3 decomposition were only present in low relative abundance in the total Illumina sequence library. For example, subgroups that become much more pronounced in the NMF data are subgroup 13 (associated with the low-pH NMF cluster) and subgroup 17 (associated with the high-pH NMF cluster).

Additional acidobacterial subgroups associated with pH reinforce the ability of NMF and large data sets to recover low-abundance OTUs with abundances that shift with soil pH. Although *Acidobacteria* subgroups 7 and 16 were observed previously to increase with increasing pH (Jones et al., 2009), our NMF analysis identified these groups as being most important in the rank-3 (medium pH) NMF cluster. The UniFrac-based PCoA plots (Fig. 3.5) also showed OTUs of *Acidobacteria* groups 7 and 16 occupying a central location of the biplot. A possible explanation for this is that organisms within acidobacterial subgroups 7 and 16 may actually be best adapted to a below neutral soil pH environment (e.g. pH 5.5–6.5). Other low-pH-associated groups identified by NMF were *Dyella* and *Rhodanobacter* (within the *Gammaproteobacteria*), with similar organisms observed in low-pH environments previously (Green et al., 2012; Lu et al., 2010).

The medium-pH rank-3 NMF cluster contained fewer OTUs and sequences than the high and low-pH clusters (Figs 3.6 and 3.8). Taxa specific to this cluster include *Anaerolineae*, which is a class within the *Chloroflexi* phylum, and had been observed before in clay loam acidic soils (Russo et al., 2012). Other groups, such as *Verrucomicrobia*, represented a higher proportion in the NMF rank-3 medium-pH cluster as well, although this group exhibits higher diversity in the high-pH NMF cluster. Sequences classifying to the genus *Lysobacter* (within *Gammaproteobacteria*) were also identified in the medium and high pH NMF clusters, but not within the low-pH cluster. Previously, the abundance of this genus was shown to correlate positively with pH (Postma et al., 2011).

Our primary observation that soil pH governs soil bacterial diversity and composition

was supported by both Illumina sequence data (Fig. 3.5) and DGGE fingerprint analysis (Fig. 3.2, Fig. 3.3). Although bacterial fingerprints were used previously to identify a strong link between bacterial diversity and pH (Fierer and Jackson 2006), here we used group-specific primers (Mühling et al., 2008) to focus on subsets of the bacterial community, which reduces the overall number of template targets and, theoretically, reduces fingerprint complexity. Overall, clustering of DGGE data using group-specific primers paralleled the Illumina sequence data, with fingerprints revealing shifts in certain groups from high to low pH soils. For example, fingerprints generated with *Acidobacteria* primers show a clear shift across pH plots (Fig. 3.2). On the other hand, Firmicutes patterns did not change as considerably as other groups across the pH gradient, a trend consistent with that observed in other studies (Lauber et al., 2009). Lauber et al., (2009) also reported clear shifts in bacterial phylum-level representation, with Actinobacteria and Bacteroidetes relative abundance increasing with pH and acidobacterial relative abundance decreasing with increasing pH. This is in contrast to our findings, where shifts relating to pH were only seen within phyla, and not when relative abundance at the phylum level was examined.

In summary, this research generated comprehensive 16S rRNA gene baseline data, demonstrating the influence of pH on soil microbial community composition. Without a clear link between phylogeny and the functional role of organisms over the pH gradient, expanding sequencing effort from the 16S rRNA gene to metagenomic approaches would help identify functional adaptations of soil communities to varying pH. Nonetheless, an important observation of this 16S rRNA gene based research was that many pH-associated OTUs were of low relative abundance and poorly connected to established taxonomies. Another

important future goal will be the design of primers specific to rare sequences correlated with individual pH-associated clusters (e.g. Lynch et al., 2012), with the purpose of obtaining longer sequences to learn more about the phylogenetic associations of these poorly characterized soil bacteria.

Chapter 4

Influence of biochar amendment on agricultural soil microbial communities

4.1 Introduction

Global climate change is one of the greatest environmental challenges of this century. A steady increase of CO_2 and other heat-trapping greenhouse gasses (CH_4, N_2O and halocarbons) in the atmosphere since the industrial revolution are a direct result of human activity (Smith et al., 2013; Solomon et al., 2009). Burning of fossil fuels, deforestation and modern agriculture have contributed to the current climate change crisis, with the mitigation of carbon emissions recognized as important for circumventing the worst impacts of climate change. As a result of cellular respiration, CO_2 is also produced naturally as a result of organic matter decay.

Soil is a major reservoir of organic carbon, estimated to be ~1,500 Pg globally (Lal 2008). However, agricultural soils are often carbon-limited, making them an ideal candidate for carbon sequestration (Hua et al., 2014). Limiting CO_2 release into the atmosphere by re-capturing carbon in soil in a recalcitrant form, such as biochar, has potential to benefit soil productivity and mitigate climate change (Mao et al., 2012; Woolf et al., 2010). It has been estimated that a 5% increase in soil organic carbon has the ability to offset atmospheric carbon by ~16% (Hua et al., 2014). Not only is the storage of biochar in soil carbon-neutral,

as the compound is very resistant to decomposition, but it also has the potential to improve soil fertility (Mao et al., 2012; Woolf et al., 2010).

Biochar is very similar to charcoal because they are both recalcitrant aromatic-carbon compounds. However, biochar is produced with the expectation of it being used as an agricultural amendment (Lehmann and Joseph 2009). Biochar is formed during low temperature pyrolysis, with temperature, duration and input materials affecting the end product characteristics. Many different organic feedstocks can be used for biochar production, ranging from lignocellulose sources to poultry waste (Azargohar et al., 2014). However, to ensure maximal carbon-neutral benefits, waste material that is not in direct competition with food production should be used. Biochar production is an exothermic reaction with bio-oil, syngas and heat also being produced. These by-products can also be used in energy generation, further offsetting the carbon foot print (Woolf et al., 2010).

The first recorded use of biochar was in pre-Columbian Amazonian Anthrosols used by ancient peoples to grow food crops. Amazonian tropical soils are known as highly weathered and nutrient poor soils that are not well suited to agriculture. In general, high precipitation and average yearly temperatures in humid tropical zones result in rapid mineralization of soil organic matter (Glaser et al., 2001). Even 500 to 9,000 years later, Amazonian biochar-amended soils are still highly recognizable, in comparison to surrounding unamended soils, due to higher carbon content, more nutrients and increased cation exchange capacity (Grossman et al., 2010; Liang et al., 2006).

Biochar is a recalcitrant and condensed aromatic compound, with a half-life estimated at 10^2 to 10^3 years. This long residency time in a soil environment ensures that carbon release

into the atmosphere is low (Gonzalez et al., 2005). As a soil amendment, biochar can have positive effects on crop yield, soil pH, nutrient retention, fertilizer requirements and productivity (Lehmann and Joseph 2009). Biochar benefits crops via its "indirect nutrient value", which increases the availability of nutrients and its nutrient-holding ability (Glaser et al., 2001). Biochar amendment has also been shown to affect soil structure, porosity, and particle size (Atkinson et al., 2010)

Microbial communities in Amazonian Anthrosols are more diverse than surrounding soil (O'Neill et al., 2009), and an increased number of cultured bacterial isolates that can be recovered from these soils. However, this study used culturing as a way to assess the soil microbiota, with the obvious caveat that only a small proportion of the microbial committees are readily cultivated (Amann and Ludwig 2000). Because of this, culture-independent methods are essential for gaining a better understanding of biochar impacts on soil microorganisms. Using molecular methods, the starting material for biochar production has also been found to influence the resulting microbial communities in soil (Steinbeiss et al., 2009). Many tropical low-nutrient soils also are acidic (Lehmann and Joseph 2009). Importantly, biochar can alter soil pH, depending on the starting pH and the biochar parent material (Lehmann et al., 2011). With bacterial diversity and composition linked to soil pH (Fierer and Jackson 2006), there is strong potential for biochar to affect the microbiota of soil, indirectly, via changes in soil pH.

Although the effect of biochar on tropical soils has been well studied, its effectiveness in temperate soils has not received sufficient attention. In order to assess the impact of biochar application on temperate agricultural soils (in terms of existence of "beneficial"

biochar associated microbial communities), two experiments were conducted, involving an agricultural field application trial and a microcosm study.

4.2 Methods

4.2.1 Field sample collection

Two studies were conducted to determine the impact of biochar application on soil microbial communities: a Canadian field trial and a controlled microcosm study. Biochar field trials were run at the agricultural farm of Macdonald Campus of McGill University. The biochar originated from pine chips and was produced at ~ 500 °C for 12 minutes. The site is located close to the Ottawa River, which both formed the valley the sites are located in and is responsible for the contrasting soil types within close proximity at this site. A loamy soil and a sandy soil were selected for this study to investigate the effect of biochar on differential water retention. Biochar was applied topically and then raked into the soil plots. Three crop types were grown on both soil types: corn, soybean, and switchgrass, not in rotation. Each sample was composed of three composite core samples taken from the top 10 cm of soil, then homogenized. Three replicate soil samples were taken from either 0 t ha^{-1} biochar or 40 t ha^{-1} biochar application rates. This was repeated over the three crop types for both bulk and rhizosphere soil and for sand and clay loam soil types (Fig. 4.1). Soils and vegetation from each location were pseudo-randomly sampled and transported on ice to the lab for processing within 24 hours. Bulk soil samples were homogenized, with stones and vegetation removed with sterile forceps. Rhizosphere soil was collected using a modified method Kirk et al., (2005). Briefly, the collected plant root samples were agitated to remove excess bulk soil,

and a sterile razor blade was used to scrape material in close proximity to the root. Triplicate subsamples from each plot (either bulk or rhizosphere samples) were combined in preparation for DNA extraction as described below.

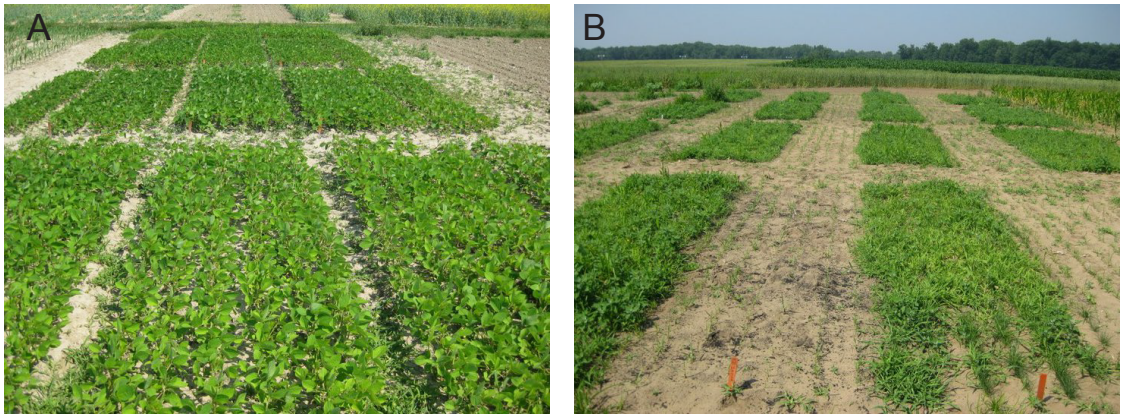

Figure 4.1 Example photographs of field study site: (A) soybean plots, (B) switchgrass plots.

4.2.2 Microcosm study sample collection

In order to study the effect of biochar on soil without the influence of vegetation, a microcosm study was conducted. The microcosm study was run over a period of twenty weeks, with biochar added at rates equivalent to 0, 20, 40, and 60 t ha^{-1} to a loam soil in a closed incubation system (1 L Mason jar). Rates of biochar application were chosen as equivalent or higher than would be used in the field in order to intensify the biochar effect, if any. Unlike the field study, the biochar used here was ground before adding to the soil. Biochar particle sizes between 2 to 20 mm have not been found to have an significant effects on crop yields (Atkinson et al., 2010). Subsamples of the soil incubation study were collected from where the study was conducted at McGill University and sent on dry ice to the University of Waterloo where the soil was processed for DNA extraction.

4.2.3 Nucleic acid extraction

DNA was extracted from all soil samples using the MoBio Power Soil DNA Isolation Kit (Carlsbad, CA) according to the manufacturer's instructions. Extracted DNA was used as template for bacterial 16S rRNA gene fingerprinting (denaturing gradient gel electrophoresis; DGGE) and amplicon sequencing (101,448,506 assembled sequences generated by Illumina). Illumina data can be found on NCBIs SRA under the accession number SUB799438.

4.2.4 DGGE-PCR

DGGE was conducted on all samples using the universal bacterial primers and reaction conditions outlined by Muyzer et al., (1993). Equal ng amounts of the resulting PCR product were loaded into each well and gels were run at 85 V for 14 h. Gels consisted of 8% acrylamide and bis-acrylamide (37.5:1), with a denaturing gradient from 40% to 60% (100% denaturant contains 7 M urea and 40% formamide). Gels were stained with SYBR Green I nucleic acid stain (Bio-Rad) and scanned on a Pharos FX Imager (Bio-Rad). Band of interest were excised, re-amplified with the original primers and Sanger sequenced.

4.2.5 Illumina library generation and sequencing.

4.2.5.1 Extracted DNA was used as template for Illumina library construction with indexed primers, as detailed in Chapter 2 of this thesis. For PCR, ~10 ng of DNA from each composite sample was added to triplicate PCR amplifications run for 20 cycles. The resulting products of these replicate reactions were pooled. Pooled amplicon templates were analyzed

by agarose gel electrophoresis and absorbance (NanoDrop; Thermo Scientific) to verify concentration and size. Paired-end sequencing (2 x 125 bases; 6-base index read) was performed on the Genome Analyzer IIx (Illumina).

4.2.6 Analysis

Paired-end reads were assembled using PANDAseq (Masella et al., 2012). This program considers base-call quality-data when determining the most likely base call over ambiguous bases or mismatches in the sequence overlap region. PCR primer sequences were also removed. AXIOME (Lynch et al., 2013) was used to manage QIIME analyses (Caporaso et al., 2010b). Sequences were clustered at 97% and representative OTUs were selected using UCLUST, with taxonomy determined using the naïve Bayesian classifier of the RDP-II (Wang et al., 2007), with a threshold cutoff of 0.5. Sequence clusters were aligned using PyNAST (Caporaso et al., 2009) and a phylogeny constructed using UniFrac (Lozupone et al., 2006; Lozupone et al., 2010). Ordinations were calculated based on weighted and unweighted UniFrac distance matrices where indicated. Indicator species analysis (Dufrene and Legendre 1997) was used to locate OTUs of interest. This method identifies species (or OTUs) that associate with a treatment of interest, with sample distributions reflecting fidelity (only in samples of that treatment) and specificity (in all samples of that treatment). This gives a number between 0 and 1, called the indicator value, and also assigns a significance to the indicator value based on a set p value (Dufrene and Legendre 1997).

4.3 Results

Taxonomic abundance profiles for the microcosm Illumina sequence data at the phylum level indicated similar overall soil taxonomic composition (Fig. 4.2). The *Acidobacteria* (17.7%), *Actinobacteria* (18.2%), *Proteobacteria* (28.5%) and unclassified taxa (25.2%) comprised the majority of sequences for each sample.

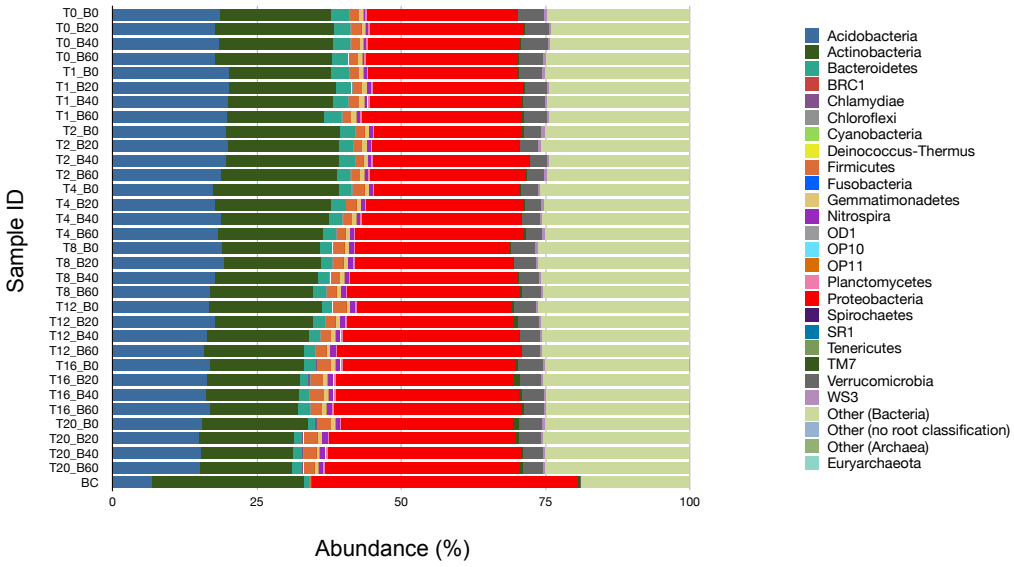

Figure 4.2 Phylum-level proportions for all microcosm samples. Within the sample ID names, T is time in weeks, and B refers to biochar treatment (0, 20, 40, or 60 t ha^{-1}). BC represents DNA extracted from biochar only.

In contrast to overall taxonomic abundances, ordination for all samples in the microcosm study demonstrated strong separation between biochar-amended and control

treatment samples (Fig. 4.3). Although biochar application influenced microcosm microbial community composition, incubation time also resulted in strong separation of samples based on microbial community composition. Ordination based on the Bray-Curtis dissimilarity metric also exhibited clear shifts in sample microbial composition based on biochar amendment (Fig. 4.4A) and also time (Fig. 4.4B). When the response variable biplot is overlayed on the plot, biochar demonstrated a strong correlation within the ordination space (Fig. 4.4C). With both methods of ordination, samples from 0-4 weeks were tightly grouped, whereas samples from 8-20 weeks exhibited more separation across Axis 2 (Fig. 4.3, 4.4).

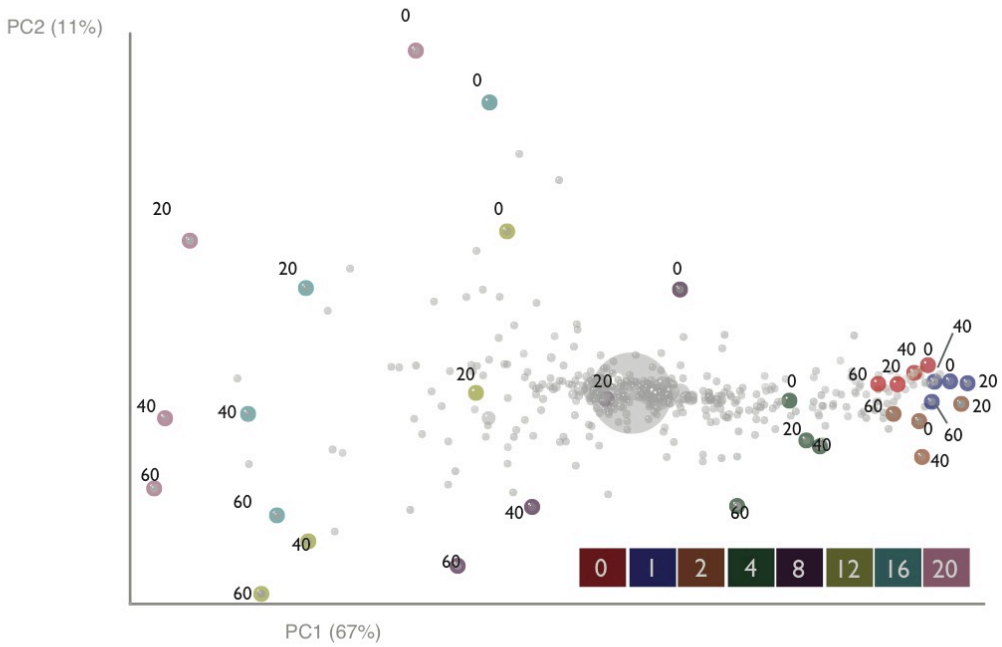

Figure 4.3 PCoA ordination of unweighted UniFrac distances for microcosm samples. Sample colours represent incubation time in weeks and numbers beside each sample represent biochar application rates (0, 20, 40 or 60 t ha^{-1}). Gray spheres represent taxonomic groups that correlate within the ordination space.

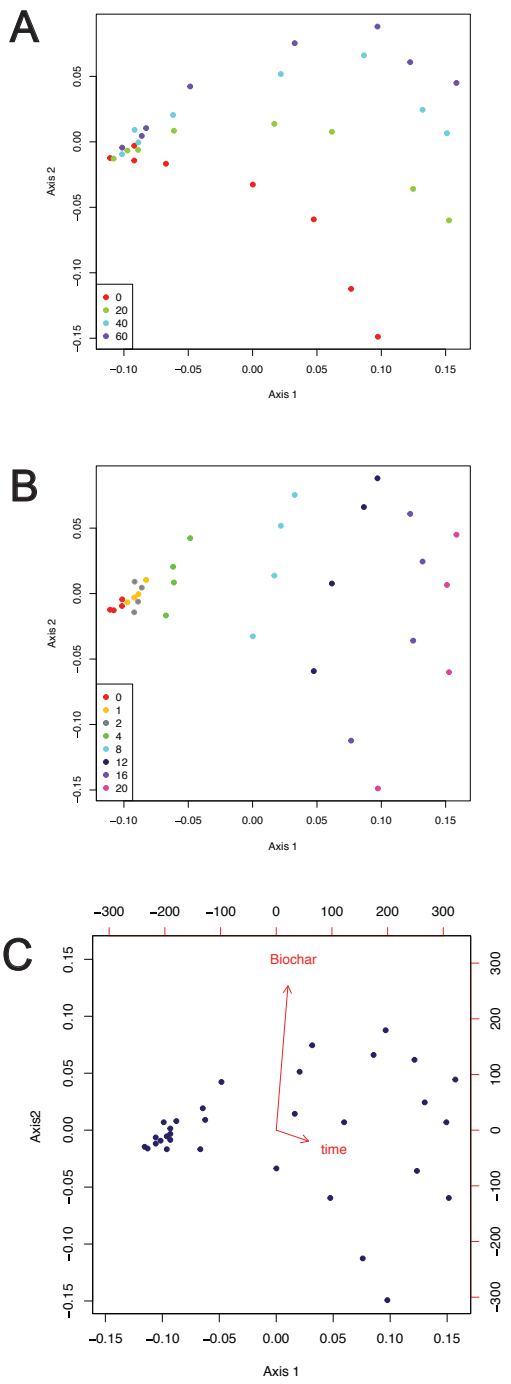

Figure 4.4 PCoA ordination for Bray-Curtis distance matrix for microcosm study samples indicating biochar application rates (0, 20, 40 or 60 t ha^{-1}; A), time (0, 1, 2, 4, 8, 12, 16 or 20 weeks; B), available sample metadata correlations within the ordination space (C).

In contrast to the microcosm experiment, the field study results suggested that the dominant factor governing agricultural soil microbial communities was soil type, followed by plant type, with no visible separation of samples based on biochar application (Fig. 4.3). The separation of soil type, sandy and loam, is very clear, with the majority of separation occurring on the second axis (Fig. 4.5A). Plant cover and biochar application had the least effect, with some slight clustering occurring among the switchgrass samples (Fig. 4.5B,C). Many of the samples pertaining to rhizospere soil seem to cluster closer to the right, indicating that contamination of rhizosphere soil with bulk soil likely occurred (Fig. 4.5D)

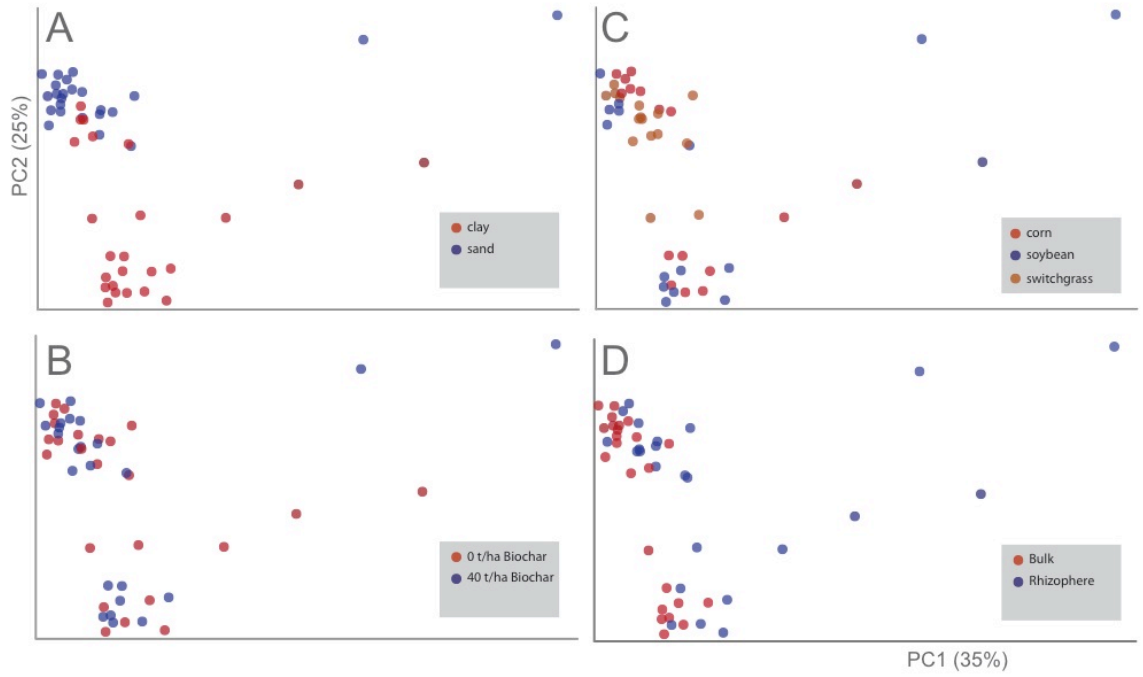

Figure 4.5 PCoA ordination based on weighted UniFrac distances for all field study samples coloured by soil type (A), biochar application rate (B), surface vegetation (C) or rhizosphere versus bulk soil (D).

The bacterial DGGE fingerprints for the microcosm study revealed a predominant biochar-associated band (Fig 4.6A; with a DDBJ accession number LC020102). The corresponding, very similar sequence, increased proportionally in biochar-amended samples and was absent from the unamended controls (Fig 4.6B). This sequence was classified as Gammaproteobacteria using the RDP-classifier. Interestingly, this sequence was present in the biochar-only sequenced sample (sample BC), suggesting that the biochar itself served to inoculate the microcosms with these bacteria, which then dominated the microcosms increasingly over time.

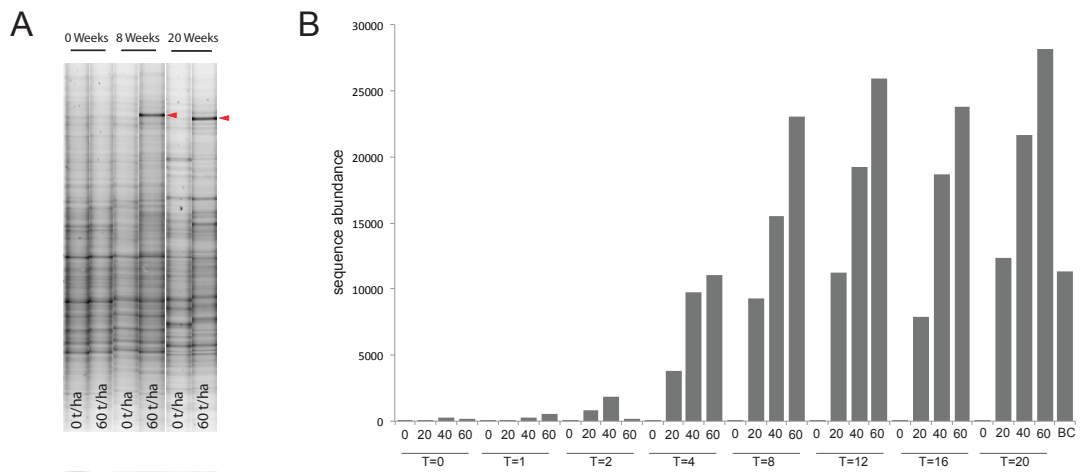

Figure 4.6 Abundance of a specific biochar-associated bacterium, showing DGGE fingerprints of bacterial 16S rRNA genes for the microcosm study 0 and 60 t ha^{-1} biochar treatments (A). The red triangles indicate a predominant band that occurs in biochar-amended samples following extended incubations. The occurrence of this sequenced band within corresponding Illumina library data revealed a proportional abundance of this sequence over time (B). The Illumina sequence had a 95% identity to the DGGE band sequence, with an e-value of $6e^{-32}$.

Indicator species analysis (Dufrene and Legendre 1997) was conducted on grouped microcosm samples from later time points (i.e. 8, 12, 16, and 20 weeks), in order to identify operational taxonomic units (OTUs) abundant in biochar application with high fidelity and specificity. In this study each biochar treatment was compared to the no-biochar control. Similar indicator species were observed for all biochar applications. For example, *Acidobacteria* subgroups 5, 6, 7, and 10 were all strong indicators of biochar application (Fig. 4.7A). Unclassified *Actinomycetales*, and OTUs from the *Rubrobacterineae* were indicators for low to medium biochar application, whereas the genus *Pseudonocardia* was

associated with the high biochar application rates (Fig. 4.7B). Within the phylum *Bacteroidetes*, unclassified *Bacteroidetes* represented the majority of indicator species, with *Flavobacteriaceae* and *Sphingobacteriales* also associated with biochar application (Fig. 4.7C). Unclassified *Gammaproteobacteria* comprised almost half of all proteobacterial indicator species, with unclassified *Alphaproteobacteria* and the genus *Enhygromyxa* (within the *Deltaproteobacteria*) also represented (Fig. 4.7D).

Indicator species analysis of field study samples revealed *Sphingobacterium* (within the phylum *Bacteroidetes*) as the most abundant indicator species, with unclassified *Betaproteobacteria*, unclassified *Gammaproteobacteria* and unclassified Bacteria also represented as indicators (Fig. 4.8).

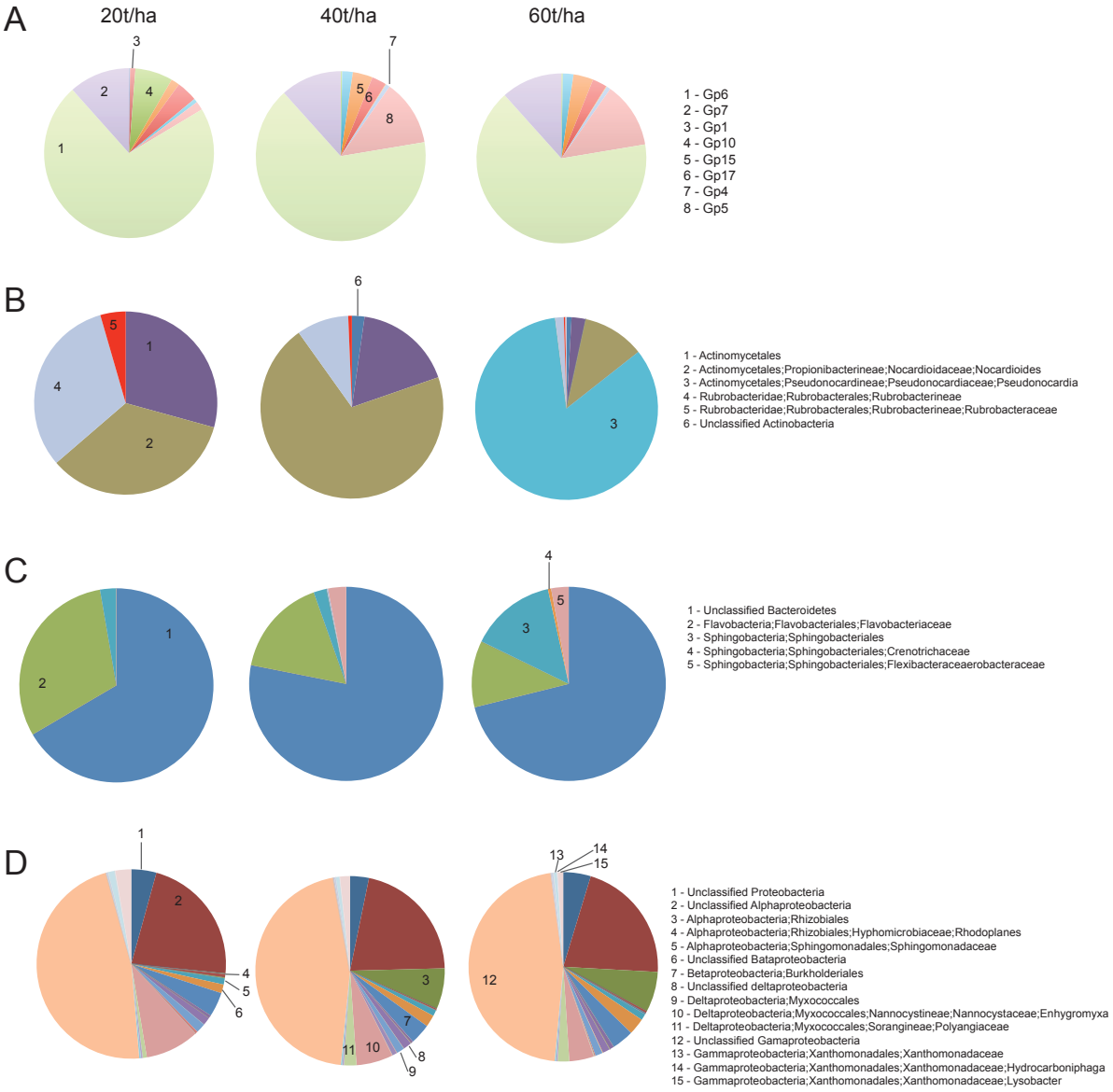

Figure 4.7 Biochar-specific indicator species (indicator values >0.7, $p < 0.05$) for the microcosm study, with biochar applications at 20, 40 and 60 t ha^{-1} and for grouped samples from 8-20 weeks. Indicator species are summarized for the *Acidobacteria* (A), *Actinobacteria* (B), *Bacteroidetes* (C), and *Proteobacteria* (D). For complete list of taxa see Appendix Table B.1, B.2 and B.3.

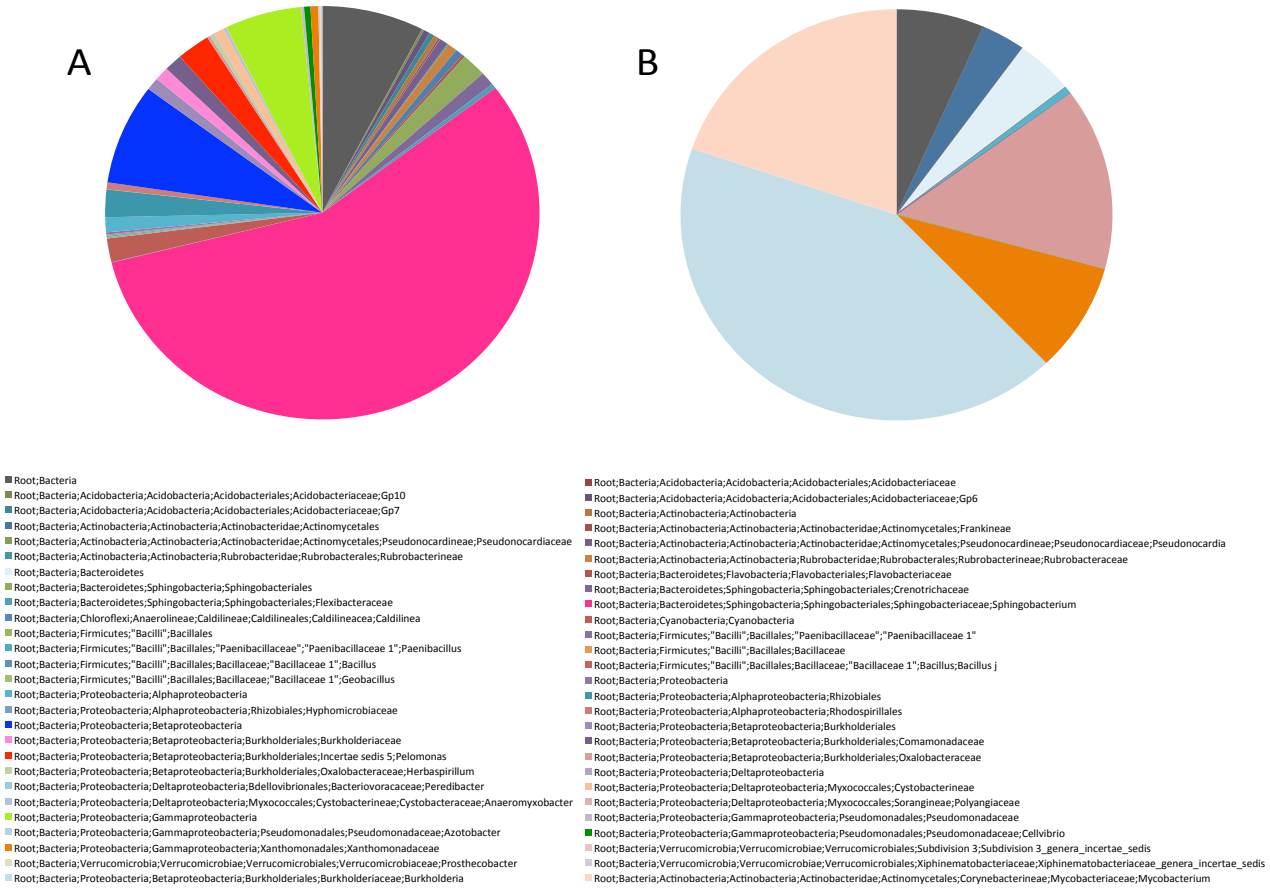

Figure 4.8 Biochar-associated indicator species (indicator values >0.7, $p < 0.05$) for the field study, (A) for all 40 t ha^{-1} treatments and (B) no biochar amendment control.

Complete biochar-associated indicator species for 8-20 week samples from the microcosm experiment and biochar field studies were clustered, revealing overlap between the two sets of indicators (Table 4.1). Common indicator species are unclassified bacterial OTUs as well as unclassified *Proteobacteria* (including unclassified *Alphaproteobacteria* and *Gammaproteobacteria*). Acidobacterial subgroups 4, 6 and 7 are common between the two studies, in addition to *Nocardioidaceae* (present as indicator species in the low and medium microcosm biochar application levels). *Nocardioidaceae* are frequently found in soil and water samples (possibly originating from a distance via an aerial source; Favet et al., 2013) with some genera considered opportunistic pathogens. Some *Nocardioidaceae* may also be hydrocarbon degraders (Luo et al., 2014). Because complete sets of indicator species, with no cut off applied were clustered, both studies contain IVs that are below 0.7. It should also be noted that higher indicator values were more common for the microcosm study than the field study.

Table 4.1. Biochar-associated indicator species associated with microcosm and field studies and their associated indicator value.

Cluster	Microcosm-IV	Field - IV	Consensus lineage
1	0.4167	0.4286	Root;Bacteria
2	0.386	0.2083	Root;Bacteria
3	0.4023	0.4189	Root;Bacteria
4	0.3906	0.4936	Root;Bacteria
5	0.9068	0.575	Root;Bacteria
6	0.8077	0.4062	Root;Bacteria

7	0.7701	0.4318	Root;Bacteria
8	0.7017	0.4167	Root;Bacteria
9	0.637	0.5514	Root;Bacteria
10	0.63	0.5098	Root;Bacteria
11	0.7076	0.5135	Root;Bacteria
12	0.896	0.6496	Root;Bacteria
13	0.5746	0.5	Root;Bacteria
14	0.8859	0.3889	Root;Bacteria
15	0.4127	0.5089	Root;Bacteria
16	0.3704	0.4464	Root;Bacteria
17	0.9561	0.516	Root;Bacteria
18	0.8578	0.4257	Root;Bacteria
19	0.8073	0.5521	Root;Bacteria
20	0.8072	0.4712	Root;Bacteria
21	0.7557	0.451	Root;Bacteria
22	0.7083	0.4808	Root;Bacteria
23	0.6917	0.4359	Root;Bacteria
24	0.6899	0.3811	Root;Bacteria
25	0.6796	0.5	Root;Bacteria
26	0.4085	0.3461	Root;Bacteria
27	0.8685	0.4091	Root;Bacteria
28	0.8477	0.4493	Root;Bacteria
29	0.6119	0.2999	Root;Bacteria
30	0.5406	0.7109	Root;Bacteria
31	0.8562	0.4861	Root;Bacteria
32	0.6412	0.5306	Root;Bacteria
33	0.3095	0.4	Root;Bacteria
34	0.7083	0.4545	Root;Bacteria
35	0.5785	0.4167	Root;Bacteria;Acidobacteria;Acidobacteria;Acidobacteriales;Acidobacteriaceae;Gp4
36	0.6762	0.375	Root;Bacteria;Acidobacteria;Acidobacteria;Acidobacteriales;Acidobacteriaceae;Gp6
37	0.9863	0.4493	Root;Bacteria;Acidobacteria;Acidobacteria;Acidobacteriales;Acidobacteriaceae;Gp7
38	0.8522	0.518	Root;Bacteria;Actinobacteria;Actinobacteria;Actinobacteridae;Actinomyc

			etales
39	0.4583	0.4018	Root;Bacteria;Actinobacteria;Actinobacteria;Actinobacteridae;Actinomycetales
40	0.6526	0.436	Root;Bacteria;Actinobacteria;Actinobacteria;Actinobacteridae;Actinomycetales;Propionibacterineae;Nocardioidaceae;Nocardioides
41	0.7185	0.3068	Root;Bacteria;Bacteroidetes;Sphingobacteria;Sphingobacteriales;Crenotrichaceae
42	0.5712	0.3409	Root;Bacteria;Firmicutes
43	0.6927	0.4375	Root;Bacteria;Firmicutes
44	0.8999	0.3846	Root;Bacteria;Firmicutes;"Bacilli";Bacillales
45	0.637	0.3977	Root;Bacteria;Firmicutes;"Bacilli";Bacillales
46	0.8681	0.4978	Root;Bacteria;Firmicutes;"Bacilli";Bacillales;Bacillaceae
47	0.5578	0.5	Root;Bacteria;Proteobacteria
48	0.9158	0.4322	Root;Bacteria;Proteobacteria
49	0.7428	0.5809	Root;Bacteria;Proteobacteria
50	0.9371	0.5485	Root;Bacteria;Proteobacteria
51	0.3723	0.4514	Root;Bacteria;Proteobacteria
52	0.9004	0.4242	Root;Bacteria;Proteobacteria;Alphaproteobacteria
53	0.8012	0.39	Root;Bacteria;Proteobacteria;Alphaproteobacteria
54	0.685	0.435	Root;Bacteria;Proteobacteria;Alphaproteobacteria
55	0.7361	0.3769	Root;Bacteria;Proteobacteria;Alphaproteobacteria;Rhizobiales
56	0.7308	0.3012	Root;Bacteria;Proteobacteria;Alphaproteobacteria;Rhodospirillales
57	0.9712	0.5642	Root;Bacteria;Proteobacteria;Betaproteobacteria
58	0.9441	0.5646	Root;Bacteria;Proteobacteria;Betaproteobacteria
59	0.7856	0.4651	Root;Bacteria;Proteobacteria;Betaproteobacteria
60	0.6269	0.6122	Root;Bacteria;Proteobacteria;Betaproteobacteria
61	0.9601	0.4694	Root;Bacteria;Proteobacteria;Betaproteobacteria;Burkholderiales
62	0.6325	0.7755	Root;Bacteria;Proteobacteria;Betaproteobacteria;Burkholderiales
63	0.9463	0.4626	Root;Bacteria;Proteobacteria;Betaproteobacteria;Burkholderiales;Burkholderiaceae
64	0.7132	0.5577	Root;Bacteria;Proteobacteria;Betaproteobacteria;Burkholderiales;Comamonadaceae
65	0.3632	0.3472	Root;Bacteria;Proteobacteria;Betaproteobacteria;Burkholderiales;Comamonadaceae
66	0.8818	0.4778	Root;Bacteria;Proteobacteria;Betaproteobacteria;Burkholderiales;Oxaloba

			cteraceae
67	0.9978	0.4706	Root;Bacteria;Proteobacteria;Gammaproteobacteria
68	0.9156	0.8152	Root;Bacteria;Proteobacteria;Gammaproteobacteria
69	0.875	0.67	Root;Bacteria;Proteobacteria;Gammaproteobacteria
70	0.7917	0.4531	Root;Bacteria;Proteobacteria;Gammaproteobacteria
71	0.742	0.5088	Root;Bacteria;Proteobacteria;Gammaproteobacteria
72	0.9003	0.4634	Root;Bacteria;Proteobacteria;Gammaproteobacteria;Pseudomonadales;Pseudomonadaceae
73	0.75	0.5385	Root;Bacteria;Proteobacteria;Gammaproteobacteria;Pseudomonadales;Pseudomonadaceae;Azotobacter

4.4 Discussion

The impact of biochar application on microbial communities within a variety of soil types has not been well characterized previously. A recent study used pyrosequencing of 16S rRNA gene amplicons to investigate the impact of biochar on Amazonian anthrosols and adjacent soils (Taketani et al., 2013). *Acidobacteria* made up a large proportion of the biochar-associated 16S rRNA genes, with subgroups 5 and 6 dominating biochar-amended soils. The authors also point out that soil formation is a stochastic process that makes it difficult to predict the outcome that follows after the addition of biochar. Another study examined microbial network associations found associations between biochar application and *Acidobacteria* Gp1 and Gp3 in addition to OTUs associated with *Actinobacteria*, *Bacillales*, and *Burkholderiaceae* from *Proteobacteria* (Nielsen et al., 2014).

Despite being characterized by similar overall phylum-level distributions (Fig. 4.2), biochar-amended microcosm samples in the current study were associated with microbial communities distinct from those in samples without biochar (Fig. 4.3, 4.4). In addition, indicator species analysis identified that specific *Acidobacteria* subgroups were associated with biochar application. For example, acidobacterial subgroups 5, 6, 7, 17, and 15 affiliated with all biochar application levels. Because of the large variability likely accountable to factors other than biochar application rates, future analysis should aim to identify the proportion of variability explained solely by biochar in comparison to other factors (such as soil type and vegetation cover). One method that would be applicable is distance-based redundancy analysis (db-RDA).

General bacterial DGGE fingerprints for the microcosm incubations further indicated that there was at least one prominent OTU that increased proportionally with higher biochar amendments and over time. This sequence was very similar to an OTU classified as *Gammaproteobacteria*, that was present in the Illumina library for the biochar-associated mesocom samples (Fig. 4.6). Additionally, the *Proteobacteria*-specific abundances indicated that there was a *Gammaproteobacteria* shift in the biochar-amended samples (Fig. 4.2). However, looking specifically at *Gammaproteobacteria* in the microcosm indicator species analysis (Fig. 4.7D), the indicator species list contains OTUs that are classified as *Gammaproteobacteria* that correspond all three biochar application levels (20, 40 and 60 t ha^{-1}; Fig. 4.7D), as well as different *Gammaproteobacteria* classified OTUs that correspond to the biochar-free sample (See Appendix Table B.1, B.2 and B.3). With multiple indicator species having consensus lineages corresponding to "unclassified *Gammaproteobacteria*", it

is difficult to interpret why specific OTUs are indicators for a specific biochar application. This may be partly a bias of using 16S rRNA gene sequencing for this analysis, as well as a limitation of available classification tools. For example, with larger reference databases there would be a reduced chance of obtaining partially identified sequences. However, poor classification of biochar-associated OTUs opens up avenues of research that can better understand these organisms and their distributions. For example, various culture-based approaches have been conducted in order gain physiological information on uncultured microorganisms (O'Neill et al., 2009).

Clustered indicator species analysis from the complete microcosm and field studies (Table 4.1) indicates that there is overlap between the two experiments, with the majority (33 or 47%) of overlapping biochar-associated OTUs affiliated with unclassified Bacteria. Three acidobacterial subgroups (4, 6, and 7) are represented in the cluster analysis. This corresponds to *Acidobacteria* subgroups associated previously with biochar. For example, *Acidobacteria* subgroup 6 was also recognized as biochar associated (Taketani et al., 2013). Biochar application has been linked to altered N_2O emission in soil (Liu et al., 2014). With *Acidobacteria* known to contain nitrite and nitrate reduction genes, the increased importance of this group of bacteria in the indicator species analysis may be linked the nitrogen cycles association with biochar (Nielsen et al., 2014). Within the *Betaproteobacteria*, the order *Burkholderiales* (with familiar groups *Burkholderiaceae*, *Comamonadaceae*, and *Oxalobacteraceae*) were also represented in the cluster comparison of indicator species. The *Burkholderiales* are known to have phosphate solubilizing genes while also being biochar

associated (Anderson et al., 2011; Nielsen et al., 2014). Biochar has been linked to increased phosphate availability, which is beneficial to plant nutrition (Anderson et al., 2011).

Because there was not an indicator value cut-off applied when assessing overlap between the two experiments, due to too few indicator species resulting when a cutoff of 0.7 was applied, many of the indicator species have low IV values. However, the clustered OTUs tend to have higher IVs when obtained from the microcosm experiment than from the corresponding field study overall, with the average IVs from the clustered subsets being 0.72 and 0.47 for the microcosm and field study, respectively. This is likely due to the field sample capturing multiple factors, not just biochar, affecting the resulting bacterial community.

Sequence library size is an important component to consider when examining the microorganism associated with biochar application. The previously discussed relationship between soil microbial communities and pH (Chapter 3) was relatively strong in regards to some taxa (specifically with some sub-groups within the Acidobacteria). However when there are many competing factors also affecting the microbial community structure, not just the one being studied, larger datasets become necessary. Because biochar application was not the only factor influencing the soil bacterial communities in the field study, other methods, such as indicator species analysis, were necessary to characterize the microbial community response.

Biochar composition and effect on the surrounding biota can vary greatly depending on the starting material, reaction temperature, and time. Because of this, further research must address whether biochar parent material and synthesis conditions influence crop growth

as well as the microbial community reponse, given that microbes play a large part in soil and crop health. Such research is essential prior to biochar's widespread adoption as an agricultural amendment and carbon sequestration option.

Chapter 5

Concluding remarks and future directions

5.1 Summary

The development of next-generation sequencing provided an opportunity to revolutionize amplicon sequencing within the context of microbial community characterization. Illumina-based sequencing provides the advantage of generating millions of reads simultaneously. Prior to the advent of Illumina, pyrosequencing had been used for microbial community amplicon sequencing (Huse et al., 2008). However, the disadvantage of this approach was a relatively high cost and lower sequence output compared to Illumina sequencing (Sbpner et al., 2011). When my thesis research started, there were no published Illumina methods available for sequencing single gene amplicons. Due to the short Illumina sequence read lengths available initially (i.e. 35 bases), my unpublished method development targeted a shorter segment of the 16S rRNA gene (i.e. the V1 hypervariable region). This region was first used in a Sanger-based high-throughput method, before 454 pyrosequencing was available, that used concatenated V1 regions that were then sequenced to generate individual ribosomal sequence tags (Neufeld et al., 2004). However, the Illumina sequencing platform quickly migrated to 125-bp reads, which enabled sequencing of the V3 region (~160 bases in length) as an alternative target.

Prior to the publication of Chapter 2's Illumina sequencing method (Bartram et al., 2011), other researchers had developed comparable techniques. For example, Lazarevic et al., (2009) sequenced the V5 region of an oral biofilm sample. Caporaso et al., (2010c)

selected the V4 region and sequenced PCR amplicons originating from various human body locations, as well as from soil, sediment, and water. In addition to Illumina sequencing of six "tandem-variable regions" from a stool sample, Claesson et al., (2010) also evaluated the regions artificially by cropping the variable regions from a full 16S rRNA gene. Although the above studies were some of the first microbial ecology studies to utilize the Illumina sequencing platform, none took advantage of assembly of paired-end reads as a method for error correction. In addition to the above-mentioned studies, and in contrast to the method developed in this thesis, other prior methods were limited to shorter Illumina sequence reads (Caporaso et al., 2010a; Gloor et al., 2010; Hummelen et al., 2010; Lazarevic et al., 2009; Zhou et al., 2010a).

Since publication of the Illumina method (Chapter 2), many microbial ecology studies have used the Illumina platform for studying microbial communities continues by accessing 16S rRNA gene amplicons. With a continued decrease in sequencing costs for this platform, and with the release of the Illumina MiSeq, deep-sequencing has largely been democratized to enable the widespread examination of additional microbial habitats by a broad range of microbial ecologists (Caporaso et al., 2012). Illumina-based sequencing methods have now been used to gain a better understanding of the human intestinal tract (Ong et al., 2013), linking vaginal microbiota with HIV infection (Hummelen et al., 2010), and disease-state human samples (Waluikar et al., 2014). Illumina amplicon sequencing has been adapted to work with many variable regions of the 16S rRNA gene as well as being coopted for the analysis of eukaryotic 18S rRNA genes (Hugerth et al., 2014).

My primary thesis objective was to develop a working, effective Illumina protocol for microbial community analysis (Chapter 2) and then apply this method for the analysis of a soil pH gradient (Chapter 3) and biochar-amended agricultural soil (Chapter 4). As shown in this thesis, with the development and implementation of high-throughput sequencing approaches, millions of sequenced partial 16S rRNA genes lend themselves well to analysis of increased sensitivity. With a known link between sample size and microbial diversity estimators' reliability (Hughes et al., 2001), increasing depth in sequence libraries increases the ability to study complex, highly diverse communities, such as soils. The sequencing method described in Chapter 2 uses primers and adaptors that can be readily modified to target other genes or increase the region of interest. Indeed, the method has now been adapted to the analysis of multiple variable regions in a single amplicon (i.e. V3-V4; Kennedy et al., 2014). In this thesis, the method was used to examine a highly diverse Arctic tundra sample (Chapter 2), identify differences in bacterial taxa over a soil pH gradient (Chapter 3) and detect shifts in microbial communities associated with the addition of biochar (Chapter 4). The results suggest that these high-throughput sequencing efforts were effective in comprehensively and reproducibly characterizing both the abundant and rare bacteria present in diverse microbial communities.

Previous studies indicated that extensive sequencing depth may be required for some environments (Fulthorpe et al., 2008). However, depending on the strength of tested treatment effects, and the extent of bacterial community diversity, a decrease in sequence library size may be acceptable for differentiating between bacterial community profiles (Kuczynski et al., 2010). For example, the effect of pH on Craibstone soil plots (Chapter 3)

was strong and would be detected even with smaller per-sample dataset sizes than those generated by my analysis. However, treatment differences related to biochar application (Chapter 4) were much weaker; larger library sizes would likely have been beneficial to better identify biochar-specific treatment effects. Although a smaller number of sequences may be sufficient to detect underlying patterns in some cases, an advantage of large sample sizes is the increased coverage of the species existing at low relative abundance (Huse et al., 2010; Sogin et al., 2006).

Despite its utility, Illumina-based 16S rRNA gene sequencing has drawbacks. Sequencing artifacts continue to be a problem, such as miscalled bases. However, assembly of paired-end reads by programs such as PANDAseq can help to mitigated these errors (Masella et al., 2012). In addition to this, because of the high sensitivity of this method, DNA contamination can be an issue in larger datasets, such as that associated with reagents or the agar used to grow pure cultures (see Chapter 2). Some artifacts can be more difficult to detect, such as those that occur during PCR itself, such as chimera formation and bias that can arise from template-primer mismatches. Kennedy et al., (2014) examined possible bias introduced by various steps associated with template concentration and PCR protocols, as well as variability in the Illumina sequencing run itself. Initial PCR template concentration was found to have an effect on downstream data analysis, whereas pooling of separate PCR amplicons was not found to affect the results significantly. Interlane sequencing differences were also not found to contribute significantly to sample profiles of either soil of fecal samples (Kennedy et al., 2014).

5.2 Future research

Novel organisms that have never been detected before, due to their low relative abundance (Pedrós-Alió 2007), can now be targeted with the increase in sequence coverage not afforded with present day high-throughput sequencing methods. These organisms may be highly novel, having escaped detection with previous sampling and analysis efforts, and may even harbor genes of potential use for industrial applications. An extension of my method development work (Chapter 2) focused on specific bacterial taxa in the Craibstone plot soils, existing at low relative abundance with unique 16S rRNA gene sequences. SSUnique (Lynch *et al*., 2012) was also used to generate a network diagram used to highlight unconnected nodes in the Alert sample sequences produced earlier (Chapter 2). Primers were designed to bind to these rare and novel sequences associated with the Arctic tundra soil (Chapter 2) and longer 16S rRNA gene sequences were recovered for these taxa (Lynch et al., 2012). This study recovered three distinct phylogenetic lineages, including an unknown branch of BRC1, a sister group to the genus *Gloeobacter*, and an highly divergent mitochondrial sequence likely associated with a phylogenetically distinct Arctic soil eukaryote. These intriguing findings suggest that this targeted approach can couple well with large 16S rRNA gene surveys associated with other large data sets (e.g. soil pH plots and biochar-associated agricultural soils) for targeting unique and low abundance taxa that correlate with aspects of soil biogeochemistry or experimental treatment. Future work should focus on examining in more detail those low abundance and phylogenetically distinct microorganisms in additional samples and environments. Eventually an important question we could ask is: are rare taxa shared across different soils or habitats?

The main focus of this thesis has been to increase our understanding of microbial community composition through various methods, with a focus on a 16S rRNA gene Illumina sequencing method. Because a complete survey from a complex environmental sample continues to escape microbial ecologists (Schloss and Handelsman 2004), new methods that attempt to characterize complex communities, such as those in the soil environment, are helping for shedding light in the dark corners of microbial habitats. Providing a more complete picture with increased sequence depth has helped to answer questions about how microbial communities are distributed, the relative abundances of microbial taxa and the overall diversity of various environments. However, there is still a disconnect between microbial taxa detected and their functions. With a large portion of the diversity of microbial communities being redundant, in terms of function (Franklin and Mills 2006), more work needs to investigate the purpose of such high diversity in ecosystem function and stability. Describing how microbial functional groups are influenced by environmental factors and drawing links between taxonomic makeup and metabolic potential are crucial next steps that will increase the predictive powers of microbial ecologists for the analysis in aquatic, terrestrial, and host-associated environments.

Appendix A

Table A-1 Nucleotide sequences of primers used in the construction of libraries for Illumina sequencing. Lowercase letters denote adapter sequences necessary for binding to the flow cell, underlined lowercase are binding sites for the Illumina sequence primers, bold uppercase highlight the index sequences (the first 12 indexes were obtained from Illumina) and regular uppercase are the V3 region primers (341F on for the forward primers and 518R for the reverse primers).

Forward Primers	Sequence (5' to 3')
V3_F	aatgatacggcgaccaccgagatctacactctttccctacacgacgctcttccgatctCCTACGGGAGGCAGCAG
V3_F modified[2]	aatgatacggcgaccaccgagatctacactctttccctacacgacgctcttccgatctNNNNCCTACGGGAGGCAGCAG
Reverse Primers	
V3_1R	caagcagaagacggcatacgagat**CGTGAT**gtgactggagttcagacgtgtgctcttcccgatctATTACCGCGGCTGCTGG
V3_2R	caagcagaagacggcatacgagat**ACATCG**gtgactggagttcagacgtgtgctcttcccgatctATTACCGCGGCTGCTGG
V3_3R	caagcagaagacggcatacgagat**GCCTAA**gtgactggagttcagacgtgtgctcttcccgatctATTACCGCGGCTGCTGG
V3_4R	caagcagaagacggcatacgagat**TGGTCA**gtgactggagttcagacgtgtgctcttcccgatctATTACCGCGGCTGCTGG
V3_5R	caagcagaagacggcatacgagat**CACTGT**gtgactggagttcagacgtgtgctcttcccgatctATTACCGCGGCTGCTGG
V3_6R	caagcagaagacggcatacgagat**ATTGGC**gtgactggagttcagacgtgtgctcttcccgatctATTACCGCGGCTGCTGG
V3_7R	caagcagaagacggcatacgagat**GATCTG**gtgactggagttcagacgtgtgctcttcccgatctATTACCGCGGCTGCTGG
V3_8R	caagcagaagacggcatacgagat**TCAAGT**gtgactggagttcagacgtgtgctcttcccgatctATTACCGCGGCTGCTGG
V3_9R	caagcagaagacggcatacgagat**CTGATC**gtgactggagttcagacgtgtgctcttcccgatctATTACCGCGGCTGCTGG
V3_10R	caagcagaagacggcatacgagat**AAGCTA**gtgactggagttcagacgtgtgctcttcccgatctATTACCGCGGCTGCTGG
V3_11R	caagcagaagacggcatacgagat**GTAGCC**gtgactggagttcagacgtgtgctcttcccgatctATTACCGCGGCTGCTGG
V3_12R	caagcagaagacggcatacgagat**TACAAG**gtgactggagttcagacgtgtgctcttcccgatctATTACCGCGGCTGCTGG
V3_13R	caagcagaagacggcatacgagat**CGTACT**gtgactggagttcagacgtgtgctcttcccgatctATTACCGCGGCTGCTGG
V3_14R	caagcagaagacggcatacgagat**GACTGA**gtgactggagttcagacgtgtgctcttcccgatctATTACCGCGGCTGCTGG
V3_15R	caagcagaagacggcatacgagat**GCTCAA**gtgactggagttcagacgtgtgctcttcccgatctATTACCGCGGCTGCTGG
V3_16R	caagcagaagacggcatacgagat**TCGCTT**gtgactggagttcagacgtgtgctcttcccgatctATTACCGCGGCTGCTGG
V3_17R	caagcagaagacggcatacgagat**TGAGGA**gtgactggagttcagacgtgtgctcttcccgatctATTACCGCGGCTGCTGG
V3_18R	caagcagaagacggcatacgagat**ACAACC**gtgactggagttcagacgtgtgctcttcccgatctATTACCGCGGCTGCTGG
V3_19R	caagcagaagacggcatacgagat**ACCTCA**gtgactggagttcagacgtgtgctcttcccgatctATTACCGCGGCTGCTGG
V3_20R	caagcagaagacggcatacgagat**ACGGTA**gtgactggagttcagacgtgtgctcttcccgatctATTACCGCGGCTGCTGG
V3_21R	caagcagaagacggcatacgagat**AGTTGG**gtgactggagttcagacgtgtgctcttcccgatctATTACCGCGGCTGCTGG
V3_22R	caagcagaagacggcatacgagat**CTCTCT**gtgactggagttcagacgtgtgctcttcccgatctATTACCGCGGCTGCTGG
V3_23R	caagcagaagacggcatacgagat**CAAGTG**gtgactggagttcagacgtgtgctcttcccgatctATTACCGCGGCTGCTGG
V3_24R	caagcagaagacggcatacgagat**CCTTGA**gtgactggagttcagacgtgtgctcttcccgatctATTACCGCGGCTGCTGG

[1] Additional 72 barcodes that would be suitable for use include: ACCACT, AGTGTC, AGAAGG, TTATCC, TTAAGG, TTCTTG, TTCAAC, TTGTGA, TTGACT, TATTCG, TATAGC, TAACTC, TACCAA, TACGTT, TAGTAC, TAGATG, TCTACA, TCTGAT, TCATGT, TGTCTA, ATTCTC, ATTGAG, ATACCT, ATGCAA, AATCCA, AATGGT, AACTAG, AACACT, AAGAGA,
ACTTAC, ACATTG, ACGAAT, AGTCAT, AGAAGT, CTTATG, CTAGAA, CATCTT,
CACATA, CCAATT, CGATTA, GTTAGT, GTAACA, GTGTAT, GATAAG, GAATCT, TTCCGT, TTCGCA, TTGGTC, TGACAG, ATCTGC, ACACGA, AGGTTC, CATGAC, GCTATC, GGACTT, GGCAAT, TCTCGG, TCAGCG, TGTGCC, TGCACG, AAGGCC, ACCAGG, AGCCTG, AGCGAC, CTACGC, CTCCAG, CCGTAG, CGGTGT, CGGAAC, GTGCTG, GAACGG, GGATGC, GGCGTA.

[2] Although not used for this study, our subsequent Illumina runs have used this modified primer. The inclusion of four maximally degenerate bases ("NNNN") maximizes the diversity during the first four bases of the run; diversity is important for identifying unique clusters. This modification allows for increased cluster density and improved base-calling accuracy.

Table A-2 Taxonomic affiliations and associated confidence values for the RDP classification of unexpected sequences within the defined community libraries (C1/C2).

Taxonomic affiliation	C1		C2	
	OTU occurrence	Average confidence value	OTU occurrence	Average confidence value
Acidobacteria	14	0.73	49	0.89
Actinobacteria	15	0.88	34	0.87
Alphaproteobacteria	30	0.82	96	0.76
Anaerolineae	1	0.92	3	0.69

Bacilli	9	0.55	4	0.32
Bacteroidetes	12	0.84	2	0.77
Betaproteobacteria	22	0.82	51	0.75
Chlamydiae			1	1.00
Chloroflexi	1	0.81	2	0.50
Clostridia	5	0.83	25	0.78
Cyanobacteria			10	0.62
Deinococci	1	0.91		
Deltaproteobacteria	6	0.80	27	0.84
Epsilonproteobacteria	3	0.67	9	0.81
Fibrobacteres			9	0.97
Flavobacteria	6	0.68	3	1.00
Fusobacteria	1	0.38		
Gammaproteobacteria	63	0.87	45	0.82
Methanococci	1	1.00		
Gemmatimonadetes			1	0.99
Nitrospira	1	0.95	1	1.00
Planctomycetacia	1	0.98	1	1.00
Sphingobacteria	21	0.87	49	0.86
Spirochaetes	1	0.99		
Verrucomicrobiae	1	0.38	17	0.87

Table A-3 Taxonomic affiliations of phyla associated with distinct abundance ranks shown in Figure 2.3b for the combined Arctic tundra library. Numbers in brackets represent the phylum proportion within the total library size for each rank (%).

Taxonomic affiliation	Abundance ranks					
	1-10	11-100	101-1000	1001-10000	10001-Doubletons	Singletons

Acidobacteria	365179 (27.4)	203500 (11.9)	332280 (14.7)	105752 (9.8)	3280 (4.5)	444 (6.8)
Actinobacteria	60444 (4.5)	344566 (20.1)	431493 (19.1)	109524 (10.1)	3689 (5.0)	509 (7.8)
Aquificae				209 (<0.05)	45 (0.1)	3 (<0.05)
Bacteroidetes	76038 (5.7)	263945 (15.4)	340140 (15.1)	137636 (12.7)	5044 (6.8)	477 (7.4)
BRC1				825 (0.1)	196 (0.3)	12 (0.2)
Chlamydiae				5920 (0.5)	1710 (2.3)	144 (2.2)
Chlorobi					11 (<0.05)	
Chloroflexi		40305 (2.4)	17990 (0.8)	11097 (1.0)	418 (0.6)	16 (0.2)
Cyanobacteria	73754 (5.5)	132774 (7.8)	67597 (3.0)	17804 (1.6)	447 (0.6)	59 (0.9)
Deinococcus-Thermus				1248 (0.1)	43 (0.1)	3 (<0.05)
Dictyoglomi				142 (<0.05)	13 (<0.05)	
Euryarchaeota					6 (<0.05)	2 (<0.05)
Fibrobacteres			1822 (0.1)	1716 (0.2)	49 (0.1)	5 (0.1)
Firmicutes			3762 (0.2)	12055 (1.1)	785 (1.1)	53 (0.8)
Fusobacteria				22 (<0.05)	3 (<0.05)	1 (<0.05)
Gemmatimonadetes		12031 (0.7)	29452 (1.3)	15172 (1.4)	532 (0.7)	39 (0.6)
Lentisphaerae				428 (<0.05)	2 (<0.05)	2 (<0.05)
Nitrospira		34141 (2.0)		975 (0.1)	36 (<0.05)	5 (0.1)
OD1				2242 (0.2)	453 (0.6)	71 (1.1)
OP10			3705 (0.2)	1563 (0.1)	80 (0.1)	2 (<0.05)
OP11				510 (<0.05)	251 (0.3)	8 (0.1)
Planctomycetes			1878 (0.1)	9743 (0.9)	882 (1.2)	121 (1.9)

Proteobacteria	756426 (56.8)	562329 (32.9)	619277 (27.4)	319814 (29.6)	32836 (44.6)	2446 (37.7)
Spirochaetes				17 (<0.05)	13 (<0.05)	
SR1				69 (<0.05)	18 (<0.05)	
TM7			50971 (2.3)	27461 (2.5)	1112 (1.5)	81 (1.2)
Verrucomicrobia		48884 (2.9)	97940 (4.3)	67884 (6.3)	3148 (4.3)	267 (4.1)
WS3			8275 (0.4)	3099 (0.3)	84 (0.1)	9 (0.1)
Unclassified		67643 (4.0)	251201 (11.1)	228913 (21.2)	18455 (25.1)	1706 (26.3)
Total	1331841	1710118	2257783	1081840	73641	6485

Table A-4 Taxonomic affiliations of classes associated with distinct abundance ranks shown in Figure 2.3b for the combined Arctic tundra library. Numbers in brackets represent the class proportion within the total library size for each rank (%).

Taxonomic affiliations	Abundance Rank					
	1-10	11-100	101-1000	1001-10000	10001-Doubletons	Singletons
"Bacilli"				2801 (0.3)	178 (0.2)	15 (0.2)
"Clostridia"			2229 (0.1)	3524 (0.3)	255 (0.3)	17 (0.3)
"Erysipelotrichi"						1 (<0.05)
Acidobacteria	365179 (27.4)	203500 (11.9)	332280 (14.7)	105752 (9.8)	3280 (4.5)	444 (6.8)
Actinobacteria	60444 (4.5)	344566 (20.1)	431493 (19.1)	109524 (10.1)	3689 (5.0)	509 (7.8)
Alphaproteobacteria	505636 (38.0)	303606 (17.8)	157245 (7.0)	35290 (3.3)	2108 (2.9)	219 (3.4)
Anaerolineae		15823 (0.9)	8916 (0.4)	2987 (0.3)	48 (0.1)	7 (0.1)
Aquificae				209 (<0.05)	45 (0.1)	3 (<0.05)
Archaeoglobi						

Bacteroidetes			1574 (0.1)	388 (<0.05)	65 (0.1)	8 (0.1)
Betaproteobacteria	72205 (5.4)	129995 (7.6)	231774 (10.3)	78076 (7.2)	3117 (4.2)	439 (6.8)
BRC1_genera_incertae_sedis				825 (0.1)	196 (0.3)	12 (0.2)
Chlamydiae				5920 (0.5)	1710 (2.3)	144 (2.2)
Chlorobia					11 (<0.05)	
Chloroflexi		24482 (1.4)	9074 (0.4)	7348 (0.7)	336 (0.5)	8 (0.1)
Chrysiogenetes						
Cyanobacteria	73754 (5.5)	132774 (7.8)	67597 (3.0)	17804 (1.6)	447 (0.6)	59 (0.9)
Deferribacteres						
Deinococci				1248 (0.1)	43 (0.1)	3 (<0.05)
Deltaproteobacteria		55955 (3.3)	68348 (3.0)	99418 (9.2)	11145 (15.1)	661 (10.2)
Dictyoglomi				142 (<0.05)	13 (<0.05)	
Epsilonproteobacteria			1058 (<0.05)	737 (0.1)	55 (0.1)	5 (0.1)
Fibrobacteres			1822 (0.1)	1716 (0.2)	49 (0.1)	5 (0.1)
Flavobacteria		26809 (1.6)	40130 (1.8)	25162 (2.3)	821 (1.1)	78 (1.2)
Fusobacteria				22 (<0.05)	3 (<0.05)	1 (<0.05)
Gammaproteobacteria	178585 (13.4)	72773 (4.3)	115415 (5.1)	66695 (6.2)	9763 (13.3)	650 (10.0)
Gemmatimonadetes		12031 (0.7)	29452 (1.3)	15172 (1.4)	532 (0.7)	39 (0.6)
Lentisphaerae				428 (<0.05)	2 (<0.05)	2 (<0.05)
Nitrospira		34141 (2.0)		975 (0.1)	36 (<0.05)	5 (0.1)
OD1_genera_incertae_sedis				2242 (0.2)	453 (0.6)	71 (1.1)
OP10_genera_incertae_sedis			3705 (0.2)	1563 (0.1)	80 (0.1)	2 (<0.05)
OP11_genera_incerta				510 (<0.05)	251 (0.3)	8 (0.1)

e_sedis						
Planctomycetacia			1878 (0.1)	9743 (0.9)	882 (1.2)	121 (1.9)
Sphingobacteria	76038 (5.7)	237136 (13.9)	257622 (11.4)	84963 (7.9)	2845 (3.9)	303 (4.7)
Spirochaetes				17 (<0.05)	13 (<0.05)	
SR1_genera_incertae_sedis				69 (<0.05)	18 (<0.05)	
TM7_genera_incertae_sedis			50971 (2.3)	27461 (2.5)	1112 (1.5)	81 (1.2)
Verrucomicrobiae		48884 (2.9)	97940 (4.3)	67884 (6.3)	3148 (4.3)	267 (4.1)
WS3_genera_incertae_sedis			8275 (0.4)	3099 (0.3)	84 (0.1)	9 (0.1)
Unclassified		67643 (4.0)	338985 (15.0)	302126 (27.9)	26808 (36.4)	2289 (35.3)

Table A- 5 Taxonomic affiliations of orders associated with distinct abundance ranks shown in Figure 2.3b for the combined Arctic tundra library. Numbers in brackets represent the order proportion within the total library size for each rank (%).

Taxonomic affiliations	Abundance Rank					
	1-10	11-100	101-1000	1001-10000	10001-Doubletons	Singletons
"Erysipelotrichales"						1 (<0.05)
"Lactobacillales"				258 (<0.05)	27 (<0.05)	4 (0.1)
"Thermoanaerobacterales"			1352 (0.1)	86 (<0.05)	6 (<0.05)	1 (<0.05)
Acidobacteriales	365179 (27.4)	203500 (11.9)	332280 (14.7)	105752 (9.8)	3280 (4.5)	444 (6.8)
Actinobacteridae		165436 (9.7)	214794 (9.5)	51873 (4.8)	1060 (1.4)	186 (2.9)
Aeromonadales				99 (<0.05)	12 (<0.05)	
Alteromonadales					5 (<0.05)	
Aquificales				209 (<0.05)	45 (0.1)	3 (<0.05)

Bacillales				2133 (0.2)	132 (0.2)	10 (0.2)
Bacteroidales			1574 (0.1)	388 (<0.05)	65 (0.1)	8 (0.1)
Bdellovibrionales			2945 (0.1)	7033 (0.7)	1012 (1.4)	35 (0.5)
Burkholderiales	72205 (5.4)	48516 (2.8)	112516 (5)	51030 (4.7)	2044 (2.8)	291 (4.5)
Caldilineae		15823 (0.9)	6123 (0.3)	2624 (0.2)	46 (0.1)	7 (0.1)
Campylobacterales				666 (0.1)	23 (<0.05)	3 (<0.05)
Caulobacterales		34564 (2)	5145 (0.2)	277 (<0.05)	88 (0.1)	3 (<0.05)
Chlamydiales				5920 (0.5)	1710 (2.3)	144 (2.2)
Chlorobiales					11 (<0.05)	
Chloroflexales			6525 (0.3)	4684 (0.4)	249 (0.3)	5 (0.1)
Chloroplast		71192 (4.2)	13631 (0.6)	7332 (0.7)	243 (0.3)	25 (0.4)
Chromatiales	178585 (13.4)	45379 (2.7)	28290 (1.3)	11079 (1)	943 (1.3)	87 (1.3)
Clostridiales			877 (<0.05)	2494 (0.2)	168 (0.2)	14 (0.2)
Coriobacteridae		22124 (1.3)	3093 (0.1)	2839 (0.3)	126 (0.2)	12 (0.2)
Deinococcales				1218 (0.1)	33 (<0.05)	3 (<0.05)
Desulfarcales				627 (0.1)	129 (0.2)	9 (0.1)
Desulfobacterales				1188 (0.1)	158 (0.2)	8 (0.1)
Desulfovibrionales			6882 (0.3)	1970 (0.2)	55 (0.1)	4 (0.1)
Desulfurellales				28 (<0.05)		
Desulfuromonales				611 (0.1)	44 (0.1)	6 (0.1)
Dictyoglomales				142 (<0.05)	13 (<0.05)	
Enterobacteriales				568 (0.1)	26 (<0.05)	4 (0.1)
FamilyI (Cyanobacteria)	73754 (5.5)		5343 (0.2)	1583 (0.1)	66 (0.1)	10 (0.2)
FamilyIII (Cyanobacteria)				117 (<0.05)	3 (<0.05)	

FamilyIV (Cyanobacteria)			9572 (0.4)		3 (<0.05)	2 (<0.05)
FamilyIX (Cyanobacteria)					9 (<0.05)	
FamilyV (Cyanobacteria)		12700 (0.7)		26 (<0.05)		1 (<0.05)
FamilyVI (Cyanobacteria)			7906 (0.4)	61 (<0.05)		1 (<0.05)
FamilyXI (Cyanobacteria)				221 (<0.05)		
FamilyXII (Cyanobacteria)			1122 (<0.05)	172 (<0.05)		
FamilyXIII (Cyanobacteria)			2863 (0.1)	692 (0.1)	21 (<0.05)	1 (<0.05)
Fibrobacterales			1822 (0.1)	1716 (0.2)	49 (0.1)	5 (0.1)
Flavobacteriales		26809 (1.6)	40130 (1.8)	25162 (2.3)	821 (1.1)	78 (1.2)
Fusobacteriales				22 (<0.05)	3 (<0.05)	1 (<0.05)
Gemmatimonadales		12031 (0.7)	29452 (1.3)	15172 (1.4)	532 (0.7)	39 (0.6)
Herpetosiphonales				1402 (0.1)	22 (<0.05)	
Hydrogenophilales				55 (<0.05)	12 (<0.05)	2 (<0.05)
Kordiimonadales						1 (<0.05)
Legionellales			3260 (0.1)	9782 (0.9)	2345 (3.2)	126 (1.9)
Lentisphaerales				428 (<0.05)	2 (<0.05)	1 (<0.05)
Methylococcales				88 (<0.05)	12 (<0.05)	2 (<0.05)
Methylophilales			3701 (0.2)	942 (0.1)	16 (<0.05)	8 (0.1)
Myxococcales			47024 (2.1)	56971 (5.3)	3814 (5.2)	275 (4.2)
Nautiliales			1058 (<0.05)	71 (<0.05)	7 (<0.05)	1 (<0.05)
Neisseriales				28 (<0.05)	37 (0.1)	2 (<0.05)
Nitrosomonadales			4005 (0.2)	792 (0.1)	65 (0.1)	3 (<0.05)
Nitrospirales		34141 (2)		975 (0.1)	36 (<0.05)	5 (0.1)
Oceanospirillales				45 (<0.05)	21 (<0.05)	1 (<0.05)
Pasteurellales				82 (<0.05)	2 (<0.05)	

Planctomycetales			1878 (0.1)	9743 (0.9)	882 (1.2)	121 (1.9)
Procabacteriales				35 (<0.05)		1 (<0.05)
Pseudomonadales			7484 (0.3)	3051 (0.3)	86 (0.1)	20 (0.3)
Rhizobiales	177246 (13.3)	199306 (11.7)	92439 (4.1)	10709 (1)	522 (0.7)	91 (1.4)
Rhodobacterales		20388 (1.2)	20617 (0.9)	2832 (0.3)	187 (0.3)	15 (0.2)
Rhodocyclales		63538 (3.7)	16005 (0.7)	9564 (0.9)	358 (0.5)	34 (0.5)
Rhodospirillales		29428 (1.7)	8598 (0.4)	6511 (0.6)	55 (0.1)	22 (0.3)
Rickettsiales			1333 (0.1)	1677 (0.2)	162 (0.2)	3 (<0.05)
Rubrobacteridae		99652 (5.8)	153621 (6.8)	37742 (3.5)	1801 (2.4)	227 (3.5)
Sphingobacteriales	76038 (5.7)	237136 (13.9)	257622 (11.4)	84963 (7.9)	2845 (3.9)	303 (4.7)
Sphingomonadales	328390 (24.7)		12991 (0.6)	716 (0.1)	104 (0.1)	21 (0.3)
Spirochaetales				17 (<0.05)	13 (<0.05)	
Syntrophobacterales				1742 (0.2)	279 (0.4)	18 (0.3)
Thiotrichales		12178 (0.7)	9763 (0.4)	2557 (0.2)	363 (0.5)	35 (0.5)
Verrucomicrobiales		48884 (2.9)	97940 (4.3)	67884 (6.3)	3148 (4.3)	267 (4.1)
Vibrionales						1 (<0.05)
Xanthomonadales		15216 (0.9)	57260 (2.5)	14110 (1.3)	422 (0.6)	56 (0.9)
Unclassified	60444 (4.5)	292177 (17.1)	626947 (27.8)	448254 (41.4)	42763 (58.1)	3368 (51.9)

Appendix B

Table B- 1 Indicator species for microcosm study associated with biochar application at 20 t ha^{-1} and for time=8, 12, 16 and 20 weeks. With indicator value >0.7 and p <0.05. Count number represent the number of OTUs identifying to a particular taxonomic affiliation and sum number is the sum of the sequence reads within all of the OTUs.

Taxonomic affiliation	20t/ha IV>0.7		0t/ha IV>70	
	Count	Sum	Count	Sum
Root	1	8	0	0
Root;Bacteria	114	22282	41	6406
Root;Bacteria;Acidobacteria;Acidobacteria;Acidobacteriales;Acidobacteriaceae	1	10	0	0
Root;Bacteria;Acidobacteria;Acidobacteria;Acidobacteriales;Acidobacteriaceae;Gp1	1	49	0	0
Root;Bacteria;Acidobacteria;Acidobacteria;Acidobacteriales;Acidobacteriaceae;Gp10	2	348	0	0
Root;Bacteria;Acidobacteria;Acidobacteria;Acidobacteriales;Acidobacteriaceae;Gp15	1	79	0	0
Root;Bacteria;Acidobacteria;Acidobacteria;Acidobacteriales;Acidobacteriaceae;Gp17	1	194	0	0
Root;Bacteria;Acidobacteria;Acidobacteria;Acidobacteriales;Acidobacteriaceae;Gp25	1	36	0	0
Root;Bacteria;Acidobacteria;Acidobacteria;Acidobacteriales;Acidobacteriaceae;Gp4	0	0	1	217
Root;Bacteria;Acidobacteria;Acidobacteria;Acidobacteriales;Acidobacteriaceae;Gp5	2	81	0	0
Root;Bacteria;Acidobacteria;Acidobacteria;Acidobacteriales;Acidobacteriaceae;Gp6	11	3498	1	11
Root;Bacteria;Acidobacteria;Acidobacteria;Acidobacteriales;Acidobacteriaceae;Gp7	3	563	1	57
Root;Bacteria;Actinobacteria;Actinobacteria	0	0	5	105
Root;Bacteria;Actinobacteria;Actinobacteria;Actinobacteridae;Actinomycetales	2	229	3	53
Root;Bacteria;Actinobacteria;Actinobacteria;Actinobacteridae;Actinomycetales;Corynebacterineae	0	0	1	20
Root;Bacteria;Actinobacteria;Actinobacteria;Actinobacteridae;Actinomycetales;Corynebacterineae;Nocardiaceae;Nocardia	0	0	1	402
Root;Bacteria;Actinobacteria;Actinobacteria;Actinobacteridae;Actinomycetales;Propionibacterineae;Nocardioidaceae;Nocardioides	1	269	0	0
Root;Bacteria;Actinobacteria;Actinobacteria;Actinobacteridae;Actinomycetales;Pseudonocardineae;Pseudonocardiaceae;Amycolatopsis	0	0	1	2193

Root;Bacteria;Actinobacteria;Actinobacteria;Actinobacteridae;Actinomycetales;Pseudonocardineae;Pseudonocardiaceae;Pseudonocardia	0	0	1	99
Root;Bacteria;Actinobacteria;Actinobacteria;Rubrobacteridae;Rubrobacterales;Rubrobacterineae	3	249	0	0
Root;Bacteria;Actinobacteria;Actinobacteria;Rubrobacteridae;Rubrobacterales;Rubrobacterineae;Rubrobacteraceae	1	35	5	210
Root;Bacteria;Actinobacteria;Actinobacteria;Rubrobacteridae;Rubrobacterales;Rubrobacterineae;Rubrobacteraceae;Conexibacter	0	0	1	17
Root;Bacteria;Bacteroidetes	7	1007	9	3690
Root;Bacteria;Bacteroidetes;Flavobacteria;Flavobacteriales;Flavobacteriaceae	2	465	0	0
Root;Bacteria;Bacteroidetes;Sphingobacteria;Sphingobacteriales	2	41	1	200
Root;Bacteria;Chlamydiae;Chlamydiae;Chlamydiales	1	62	0	0
Root;Bacteria;Chloroflexi;Chloroflexi	2	318	0	0
Root;Bacteria;Chloroflexi;Chloroflexi;Chloroflexales;Chloroflexaceae;Roseiflexus	1	883	0	0
Root;Bacteria;Cyanobacteria;Cyanobacteria;Chloroplast;Chlorophyta	0	0	1	12
Root;Bacteria;Cyanobacteria;Cyanobacteria;Chloroplast;Streptophyta	0	0	1	77
Root;Bacteria;Firmicutes	0	0	1	6
Root;Bacteria;Firmicutes;"Bacilli"	0	0	2	26
Root;Bacteria;Firmicutes;"Bacilli";Bacillales	1	19	0	0
Root;Bacteria;Firmicutes;"Bacilli";Bacillales;"Alicyclobacillaceae";Alicyclobacillus	0	0	1	35
Root;Bacteria;Firmicutes;"Bacilli";Bacillales;"Paenibacillaceae";"Paenibacillaceae 1"	1	26	0	0
Root;Bacteria;Firmicutes;"Bacilli";Bacillales;"Paenibacillaceae";"Paenibacillaceae 1";Paenibacillus	1	19	0	0
Root;Bacteria;Firmicutes;"Bacilli";Bacillales;Bacillaceae;"Bacillaceae 1";Bacillus	1	61	1	9
Root;Bacteria;Firmicutes;"Bacilli";Bacillales;Bacillaceae;"Bacillaceae 1";Bacillus;Bacillus c	1	13	0	0
Root;Bacteria;Gemmatimonadetes;Gemmatimonadetes;Gemmatimonadales;Gemmatimonadaceae;Gemmatimonas	1	26	0	0
Root;Bacteria;Nitrospira;Nitrospira;Nitrospirales;Nitrospiraceae;Nitrospira	1	22	0	0
Root;Bacteria;Planctomycetes;Planctomycetacia;Planctomycetales;Planctomycetaceae	2	52	3	92
Root;Bacteria;Planctomycetes;Planctomycetacia;Planctomycetales;Planctomycetaceae;Isosphaera	0	0	0	0
Root;Bacteria;Proteobacteria	12	3805	11	585
Root;Bacteria;Proteobacteria;Alphaproteobacteria	5	19771	4	2159
Root;Bacteria;Proteobacteria;Alphaproteobacteria;Rhizobiales	2	113	1	50
Root;Bacteria;Proteobacteria;Alphaproteobacteria;Rhizobiales;Bradyrhizobiaceae	0	0	2	21
Root;Bacteria;Proteobacteria;Alphaproteobacteria;Rhizobiales;Hyphomicrobiaceae;Hyphomicrobium	1	12	0	0
Root;Bacteria;Proteobacteria;Alphaproteobacteria;Rhizobiales;Hyphomicrobiaceae;Rhodoplanes	1	503	0	0
Root;Bacteria;Proteobacteria;Alphaproteobacteria;Rickettsiales;Incertae sedis 4;Caedibacter	1	126	0	0
Root;Bacteria;Proteobacteria;Alphaproteobacteria;Rickettsiales;Incertae sedis 4;Odyssella	0	0	0	0
Root;Bacteria;Proteobacteria;Alphaproteobacteria;Sphingomonadales;Sphingomonadaceae	1	944	0	0
Root;Bacteria;Proteobacteria;Betaproteobacteria	5	1321	3	70

Root;Bacteria;Proteobacteria;Betaproteobacteria;Burkholderiales	4	3703	1	11
Root;Bacteria;Proteobacteria;Betaproteobacteria;Burkholderiales;Comamonadaceae	2	467	0	0
Root;Bacteria;Proteobacteria;Betaproteobacteria;Burkholderiales;Comamonadaceae;Ramlibacter	0	0	1	44
Root;Bacteria;Proteobacteria;Betaproteobacteria;Burkholderiales;Incertae sedis 5	0	0	1	134
Root;Bacteria;Proteobacteria;Betaproteobacteria;Burkholderiales;Oxalobacteraceae	0	0	3	555
Root;Bacteria;Proteobacteria;Betaproteobacteria;Burkholderiales;Oxalobacteraceae;Herbaspirillum	0	0	1	23
Root;Bacteria;Proteobacteria;Deltaproteobacteria	9	1200	2	179
Root;Bacteria;Proteobacteria;Deltaproteobacteria;Bdellovibrionales;Bdellovibrionaceae;Bdellovibrio	1	21	0	0
Root;Bacteria;Proteobacteria;Deltaproteobacteria;Myxococcales	9	1440	0	0
Root;Bacteria;Proteobacteria;Deltaproteobacteria;Myxococcales;Cystobacterineae	2	378	3	102
Root;Bacteria;Proteobacteria;Deltaproteobacteria;Myxococcales;Cystobacterineae;Myxococcaceae;Myxococcus	0	0	1	15
Root;Bacteria;Proteobacteria;Deltaproteobacteria;Myxococcales;Nannocystineae;Nannocystaceae	1	23	0	0
Root;Bacteria;Proteobacteria;Deltaproteobacteria;Myxococcales;Nannocystineae;Nannocystaceae;Enhygromyxa	2	8175	0	0
Root;Bacteria;Proteobacteria;Deltaproteobacteria;Myxococcales;Sorangineae;Polyangiaceae	2	627	1	47
Root;Bacteria;Proteobacteria;Deltaproteobacteria;Myxococcales;Sorangineae;Polyangiaceae;Polyangium	3	264	5	686
Root;Bacteria;Proteobacteria;Gammaproteobacteria	7	41764	13	507
Root;Bacteria;Proteobacteria;Gammaproteobacteria;Legionellales;Coxiellaceae;Aquicella	0	0	2	53
Root;Bacteria;Proteobacteria;Gammaproteobacteria;Pseudomonadales;Pseudomonadaceae	1	266	0	0
Root;Bacteria;Proteobacteria;Gammaproteobacteria;Pseudomonadales;Pseudomonadaceae;Pseudomonas	1	30	0	0
Root;Bacteria;Proteobacteria;Gammaproteobacteria;Xanthomonadales;Xanthomonadaceae	4	1010	0	0
Root;Bacteria;Proteobacteria;Gammaproteobacteria;Xanthomonadales;Xanthomonadaceae;Dyella	0	0	1	45
Root;Bacteria;Proteobacteria;Gammaproteobacteria;Xanthomonadales;Xanthomonadaceae;Hydrocarboniphaga	1	62	0	0
Root;Bacteria;Proteobacteria;Gammaproteobacteria;Xanthomonadales;Xanthomonadaceae;Lysobacter	1	2406	0	0
Root;Bacteria;Spirochaetes;Spirochaetes;Spirochaetales;Spirochaetaceae;Spirochaeta	0	0	0	0
Root;Bacteria;TM7;TM7_genera_incertae_sedis	1	246	0	0
Root;Bacteria;Verrucomicrobia;Verrucomicrobiae;Verrucomicrobiales;Opitutaceae;Opitutus	1	40	0	0
Root;Bacteria;Verrucomicrobia;Verrucomicrobiae;Verrucomicrobiales;Subdivision 3;Subdivision 3_genera_incertae_sedis	5	806	4	743
Root;Bacteria;Verrucomicrobia;Verrucomicrobiae;Verrucomicrobiales;Verrucomicrobiaceae	1	43	0	0

Root;Bacteria;Verrucomicrobia;Verrucomicrobiae;Verrucomicrobiales;Xiphinematobacteriaceae;Xiphinematobacteriaceae_genera_incertae_sedis	1	7	0	0

Table B- 2 Indicator species for mesocom study associated with biochar application at 40 t ha^{-1} and for time=8, 12, 16 and 20 weeks. With IS value >0.7 and p <0.05. Count number represent the number of OTUs identifying to a particular taxonomic affiliation and sum number is the sum of the sequence reads within all of the OTUs.

Taxonomic affiliation	40t/ha IS>70		0t/ha IS>70	
	Count	Sum	Count	Sum
Root	3	39	0	0
Root;Bacteria	205	46442	91	26800
Root;Bacteria;Acidobacteria;Acidobacteria;Acidobacteriales;Acidobacteriaceae	0	0	1	53
Root;Bacteria;Acidobacteria;Acidobacteria;Acidobacteriales;Acidobacteriaceae;Gp1	0	0	1	207
Root;Bacteria;Acidobacteria;Acidobacteria;Acidobacteriales;Acidobacteriaceae;Gp10	1	23	0	0
Root;Bacteria;Acidobacteria;Acidobacteria;Acidobacteriales;Acidobacteriaceae;Gp13	1	168	1	48
Root;Bacteria;Acidobacteria;Acidobacteria;Acidobacteriales;Acidobacteriaceae;Gp15	3	316	0	0
Root;Bacteria;Acidobacteria;Acidobacteria;Acidobacteriales;Acidobacteriaceae;Gp16	0	0	0	0
Root;Bacteria;Acidobacteria;Acidobacteria;Acidobacteriales;Acidobacteriaceae;Gp17	4	230	0	0
Root;Bacteria;Acidobacteria;Acidobacteria;Acidobacteriales;Acidobacteriaceae;Gp2	0	0	2	1859
Root;Bacteria;Acidobacteria;Acidobacteria;Acidobacteriales;Acidobacteriaceae;Gp25	0	0	1	8
Root;Bacteria;Acidobacteria;Acidobacteria;Acidobacteriales;Acidobacteriaceae;Gp3	1	10	0	0
Root;Bacteria;Acidobacteria;Acidobacteria;Acidobacteriales;Acidobacteriaceae;Gp4	1	65	1	87
Root;Bacteria;Acidobacteria;Acidobacteria;Acidobacteriales;Acidobacteriaceae;Gp5	5	1050	0	0
Root;Bacteria;Acidobacteria;Acidobacteria;Acidobacteriales;Acidobacteriaceae;Gp6	24	5495	1	226
Root;Bacteria;Acidobacteria;Acidobacteria;Acidobacteriales;Acidobacteriaceae;Gp7	8	969	0	0
Root;Bacteria;Actinobacteria;Actinobacteria	1	52	2	54
Root;Bacteria;Actinobacteria;Actinobacteria;Actinobacteridae;Actinomycetales	7	407	7	172
Root;Bacteria;Actinobacteria;Actinobacteria;Actinobacteridae;Actinomycetales;Micromonosporineae;Micromonosporaceae	0	0	1	13
Root;Bacteria;Actinobacteria;Actinobacteria;Actinobacteridae;Actinomycetales;Prop	2	1641	0	0

ionibacterineae;Nocardioidaceae;Nocardioides				
Root;Bacteria;Actinobacteria;Actinobacteria;Actinobacteridae;Actinomycetales;Pseudonocardineae;Pseudonocardiaceae	0	0	1	195
Root;Bacteria;Actinobacteria;Actinobacteria;Rubrobacteridae;Rubrobacterales;Rubrobacterineae	1	215	0	0
Root;Bacteria;Actinobacteria;Actinobacteria;Rubrobacteridae;Rubrobacterales;Rubrobacterineae;Rubrobacteraceae	1	15	1	97
Root;Bacteria;Bacteroidetes	14	2719	12	3372
Root;Bacteria;Bacteroidetes;Flavobacteria;Flavobacteriales	0	0	1	62
Root;Bacteria;Bacteroidetes;Flavobacteria;Flavobacteriales;Flavobacteriaceae	1	573	1	17
Root;Bacteria;Bacteroidetes;Sphingobacteria;Sphingobacteriales	3	77	1	181
Root;Bacteria;Bacteroidetes;Sphingobacteria;Sphingobacteriales;Crenotrichaceae;Chitinophaga	1	7	0	0
Root;Bacteria;Bacteroidetes;Sphingobacteria;Sphingobacteriales;Flexibacteraceae	2	105	0	0
Root;Bacteria;Chlamydiae;Chlamydiae;Chlamydiales	1	50	0	0
Root;Bacteria;Chlamydiae;Chlamydiae;Chlamydiales;Parachlamydiaceae	0	0	1	64
Root;Bacteria;Chlamydiae;Chlamydiae;Chlamydiales;Parachlamydiaceae;Parachlamydia	0	0	0	0
Root;Bacteria;Chloroflexi;Chloroflexi	2	354	0	0
Root;Bacteria;Chloroflexi;Chloroflexi;Chloroflexales;Chloroflexaceae;Roseiflexus	4	3764	0	0
Root;Bacteria;Chloroflexi;Chloroflexi;Herpetosiphonales;Herpetosiphonaceae;Herpetosiphon	2	504	0	0
Root;Bacteria;Firmicutes	0	0	1	97
Root;Bacteria;Firmicutes;"Bacilli";Bacillales	2	20	0	0
Root;Bacteria;Firmicutes;"Bacilli";Bacillales;"Alicyclobacillaceae";Alicyclobacillus	0	0	1	36
Root;Bacteria;Firmicutes;"Bacilli";Bacillales;"Paenibacillaceae";"Paenibacillaceae 1";Paenibacillus	0	0	2	82
Root;Bacteria;Firmicutes;"Bacilli";Bacillales;Bacillaceae	0	0	1	10
Root;Bacteria;Firmicutes;"Bacilli";Bacillales;Bacillaceae;"Bacillaceae 1";Bacillus	1	86	1	20
Root;Bacteria;Firmicutes;"Clostridia";Clostridiales;Incertae Sedis XVIII;Symbiobacterium	0	0	1	89
Root;Bacteria;Gemmatimonadetes;Gemmatimonadetes;Gemmatimonadales;Gemmatimonadaceae;Gemmatimonas	0	0	1	10
Root;Bacteria;OD1;OD1_genera_incertae_sedis	1	25	0	0
Root;Bacteria;Planctomycetes;Planctomycetacia;Planctomycetales;Planctomycetaceae	0	0	2	78
Root;Bacteria;Proteobacteria	29	5613	24	10555
Root;Bacteria;Proteobacteria;Alphaproteobacteria	8	37291	3	1671
Root;Bacteria;Proteobacteria;Alphaproteobacteria;Rhizobiales	5	12206	0	0
Root;Bacteria;Proteobacteria;Alphaproteobacteria;Rhizobiales;Hyphomicrobiaceae;Rhodoplanes	1	404	0	0
Root;Bacteria;Proteobacteria;Alphaproteobacteria;Rhizobiales;Methylobacteriaceae;Microvirga	1	33	0	0
Root;Bacteria;Proteobacteria;Alphaproteobacteria;Rhodobacterales;Rhodobacteraceae;Amaricoccus	1	39	0	0
Root;Bacteria;Proteobacteria;Alphaproteobacteria;Rhodospirillales	1	13	0	0
Root;Bacteria;Proteobacteria;Alphaproteobacteria;Rickettsiales;Incertae sedis 4;Caedibacter	1	47	0	0

Root;Bacteria;Proteobacteria;Alphaproteobacteria;Rickettsiales;Incertae sedis 4;Odyssella	0	0	1	589
Root;Bacteria;Proteobacteria;Alphaproteobacteria;Sphingomonadales;Sphingomonadaceae	2	1770	1	8
Root;Bacteria;Proteobacteria;Betaproteobacteria	8	3538	2	238
Root;Bacteria;Proteobacteria;Betaproteobacteria;Burkholderiales	6	6106	0	0
Root;Bacteria;Proteobacteria;Betaproteobacteria;Burkholderiales;Burkholderiaceae	1	20	0	0
Root;Bacteria;Proteobacteria;Betaproteobacteria;Burkholderiales;Burkholderiaceae;Paucimonas	1	13	0	0
Root;Bacteria;Proteobacteria;Betaproteobacteria;Burkholderiales;Comamonadaceae	3	669	0	0
Root;Bacteria;Proteobacteria;Betaproteobacteria;Burkholderiales;Oxalobacteraceae	1	20	3	484
Root;Bacteria;Proteobacteria;Betaproteobacteria;Burkholderiales;Oxalobacteraceae;Herbaspirillum	0	0	1	182
Root;Bacteria;Proteobacteria;Betaproteobacteria;Burkholderiales;Oxalobacteraceae;Herminiimonas	1	103	0	0
Root;Bacteria;Proteobacteria;Deltaproteobacteria	17	2502	9	1761
Root;Bacteria;Proteobacteria;Deltaproteobacteria;Bdellovibrionales;Bdellovibrionaceae;Bdellovibrio	1	88	0	0
Root;Bacteria;Proteobacteria;Deltaproteobacteria;Myxococcales	12	2336	1	96
Root;Bacteria;Proteobacteria;Deltaproteobacteria;Myxococcales;Cystobacterineae	1	148	3	147
Root;Bacteria;Proteobacteria;Deltaproteobacteria;Myxococcales;Cystobacterineae;Cystobacteraceae;Anaeromyxobacter	3	1359	1	428
Root;Bacteria;Proteobacteria;Deltaproteobacteria;Myxococcales;Cystobacterineae;Myxococcaceae;Myxococcus	0	0	1	15
Root;Bacteria;Proteobacteria;Deltaproteobacteria;Myxococcales;Nannocystineae;Haliangiaceae;Haliangium	1	86	0	0
Root;Bacteria;Proteobacteria;Deltaproteobacteria;Myxococcales;Nannocystineae;Nannocystaceae	2	15	0	0
Root;Bacteria;Proteobacteria;Deltaproteobacteria;Myxococcales;Nannocystineae;Nannocystaceae;Enhygromyxa	4	10926	0	0
Root;Bacteria;Proteobacteria;Deltaproteobacteria;Myxococcales;Sorangineae;Polyangiaceae	6	3705	4	1230
Root;Bacteria;Proteobacteria;Deltaproteobacteria;Myxococcales;Sorangineae;Polyangiaceae;Polyangium	4	467	5	574
Root;Bacteria;Proteobacteria;Gammaproteobacteria	21	79506	25	5901
Root;Bacteria;Proteobacteria;Gammaproteobacteria;Legionellales;Coxiellaceae	0	0	1	20
Root;Bacteria;Proteobacteria;Gammaproteobacteria;Legionellales;Coxiellaceae;Aquicella	0	0	2	115
Root;Bacteria;Proteobacteria;Gammaproteobacteria;Legionellales;Legionellaceae	1	302	3	275
Root;Bacteria;Proteobacteria;Gammaproteobacteria;Pseudomonadales;Pseudomonadaceae	1	469	0	0
Root;Bacteria;Proteobacteria;Gammaproteobacteria;Pseudomonadales;Pseudomonadaceae;Pseudomonas	1	36	0	0
Root;Bacteria;Proteobacteria;Gammaproteobacteria;Xanthomonadales;Xanthomonadaceae	4	1097	1	2527
Root;Bacteria;Proteobacteria;Gammaproteobacteria;Xanthomonadales;Xanthomonadaceae;Hydrocarboniphaga	1	114	0	0

Root;Bacteria;Proteobacteria;Gammaproteobacteria;Xanthomonadales;Xanthomonadaceae;Lysobacter	2	2990	0	0
Root;Bacteria;Spirochaetes;Spirochaetes;Spirochaetales;Spirochaetaceae;Spirochaeta	0	0	1	374
Root;Bacteria;TM7;TM7_genera_incertae_sedis	1	20	0	0
Root;Bacteria;Verrucomicrobia;Verrucomicrobiae;Verrucomicrobiales	2	121	0	0
Root;Bacteria;Verrucomicrobia;Verrucomicrobiae;Verrucomicrobiales;Subdivision 3;Subdivision 3_genera_incertae_sedis	7	884	8	2145
Root;Bacteria;Verrucomicrobia;Verrucomicrobiae;Verrucomicrobiales;Verrucomicrobiaceae	2	569	0	0
Root;Bacteria;Verrucomicrobia;Verrucomicrobiae;Verrucomicrobiales;Xiphinematobacteriaceae;Xiphinematobacteriaceae_genera_incertae_sedis	0	0	3	332

Table B- 3 Indicator species for microcosm study associated with biochar application at 60 t ha^{-1} and for time=8, 12, 16 and 20 weeks. With IS value >0.7 and p <0.05. Count number represent the number of OTUs identifying to a particular taxonomic affiliation and sum number is the sum of the sequence reads within all of the OTUs.

Taxonomic affiliation	60t/ha IS>70		0t/ha IS>70	
	Count	Sum	Count	Sum
Root	3	109	3	586
Root;Bacteria	294	81971	122	42979
Root;Bacteria;Acidobacteria;Acidobacteria;Acidobacteriales;Acidobacteriaceae	0	0	1	5
Root;Bacteria;Acidobacteria;Acidobacteria;Acidobacteriales;Acidobacteriaceae;Gp1	0	0	3	448
Root;Bacteria;Acidobacteria;Acidobacteria;Acidobacteriales;Acidobacteriaceae;Gp13	0	0	1	49
Root;Bacteria;Acidobacteria;Acidobacteria;Acidobacteriales;Acidobacteriaceae;Gp15	2	292	0	0
Root;Bacteria;Acidobacteria;Acidobacteria;Acidobacteriales;Acidobacteriaceae;Gp17	5	734	2	282
Root;Bacteria;Acidobacteria;Acidobacteria;Acidobacteriales;Acidobacteriaceae;Gp2	0	0	2	5335
Root;Bacteria;Acidobacteria;Acidobacteria;Acidobacteriales;Acidobacteriaceae;Gp3	0	0	1	4
Root;Bacteria;Acidobacteria;Acidobacteria;Acidobacteriales;Acidobacteriaceae;Gp4	0	0	1	106
Root;Bacteria;Acidobacteria;Acidobacteria;Acidobacteriales;Acidobacteriaceae;Gp5	5	1219	0	0

Root;Bacteria;Acidobacteria;Acidobacteria;Acidobacteriales;Acidobacteriaceae;Gp6	30	6154	3	2412
Root;Bacteria;Acidobacteria;Acidobacteria;Acidobacteriales;Acidobacteriaceae;Gp7	16	6507	0	0
Root;Bacteria;Actinobacteria;Actinobacteria	4	176	3	130
Root;Bacteria;Actinobacteria;Actinobacteria;Actinobacteridae;Actinomycetales	5	500	9	600
Root;Bacteria;Actinobacteria;Actinobacteria;Actinobacteridae;Actinomycetales;Propionibacterineae;Nocardioidaceae;Nocardioides	2	2206	0	0
Root;Bacteria;Actinobacteria;Actinobacteria;Actinobacteridae;Actinomycetales;Pseudonocardineae;Pseudonocardiaceae;Pseudonocardia	1	16759	0	0
Root;Bacteria;Actinobacteria;Actinobacteria;Actinobacteridae;Actinomycetales;Streptosporangineae;Thermomonosporaceae;Actinomadura	0	0	1	454
Root;Bacteria;Actinobacteria;Actinobacteria;Rubrobacteridae;Rubrobacterales;Rubrobacterineae	1	306	0	0
Root;Bacteria;Actinobacteria;Actinobacteria;Rubrobacteridae;Rubrobacterales;Rubrobacterineae;Rubrobacteraceae	2	55	1	21
Root;Bacteria;Actinobacteria;Actinobacteria;Rubrobacteridae;Rubrobacterales;Rubrobacterineae;Rubrobacteraceae;Conexibacter	0	0	1	15
Root;Bacteria;Actinobacteria;Actinobacteria;Rubrobacteridae;Rubrobacterales;Rubrobacterineae;Rubrobacteraceae;Thermoleophilum	1	42	0	0
Root;Bacteria;Bacteroidetes	20	3855	14	2777
Root;Bacteria;Bacteroidetes;Flavobacteria;Flavobacteriales	0	0	2	136
Root;Bacteria;Bacteroidetes;Flavobacteria;Flavobacteriales;Flavobacteriaceae	4	602	0	0
Root;Bacteria;Bacteroidetes;Flavobacteria;Flavobacteriales;Flavobacteriaceae;Flavobacterium	0	0	1	8
Root;Bacteria;Bacteroidetes;Sphingobacteria;Sphingobacteriales	1	777	3	1019
Root;Bacteria;Bacteroidetes;Sphingobacteria;Sphingobacteriales;Crenotrichaceae	2	29	0	0
Root;Bacteria;Bacteroidetes;Sphingobacteria;Sphingobacteriales;Crenotrichaceae;Chitinophaga	0	0	1	23
Root;Bacteria;Bacteroidetes;Sphingobacteria;Sphingobacteriales;Flexibacteraceae	1	158	1	640
Root;Bacteria;Chlamydiae;Chlamydiae;Chlamydiales	0	0	1	13
Root;Bacteria;Chlamydiae;Chlamydiae;Chlamydiales;Parachlamydiaceae;Neochlamydia	0	0	1	53
Root;Bacteria;Chlamydiae;Chlamydiae;Chlamydiales;Parachlamydiaceae;Parachlamydia	1	98	0	0
Root;Bacteria;Chloroflexi;Chloroflexi	1	358	0	0
Root;Bacteria;Chloroflexi;Chloroflexi;Herpetosiphonales;Herpetosiphonaceae;Herpetosiphon	2	488	0	0
Root;Bacteria;Cyanobacteria;Cyanobacteria	0	0	1	142
Root;Bacteria;Cyanobacteria;Cyanobacteria;Chloroplast	0	0	1	9
Root;Bacteria;Firmicutes	1	6	2	225
Root;Bacteria;Firmicutes;"Bacilli";Bacillales;"Paenibacillaceae";"Paenibacillaceae 1";Cohnella	0	0	1	49
Root;Bacteria;Firmicutes;"Bacilli";Bacillales;"Paenibacillaceae";"Paenibacillaceae 1";Paenibacillus	1	45	1	43
Root;Bacteria;Firmicutes;"Bacilli";Bacillales;"Thermoactinomycetaceae";Planifilum	0	0	1	22
Root;Bacteria;Firmicutes;"Bacilli";Bacillales;Bacillaceae	1	21	1	6
Root;Bacteria;Firmicutes;"Bacilli";Bacillales;Bacillaceae;"Bacillaceae 1";Bacillus	1	94	0	0
Root;Bacteria;Firmicutes;"Clostridia";Clostridiales;"Lachnospiraceae"	0	0	1	34

Root;Bacteria;OD1;OD1_genera_incertae_sedis	0	0	1	28
Root;Bacteria;Planctomycetes;Planctomycetacia;Planctomycetales;Planctomycetaceae	0	0	1	9
Root;Bacteria;Proteobacteria	45	11202	36	14979
Root;Bacteria;Proteobacteria;Alphaproteobacteria	12	50390	6	5508
Root;Bacteria;Proteobacteria;Alphaproteobacteria;Rhizobiales	5	16424	2	3499
Root;Bacteria;Proteobacteria;Alphaproteobacteria;Rhizobiales;Hyphomicrobiaceae;Rhodoplanes	1	859	0	0
Root;Bacteria;Proteobacteria;Alphaproteobacteria;Rhodospirillales;Acetobacteraceae;Stella	1	481	0	0
Root;Bacteria;Proteobacteria;Alphaproteobacteria;Rhodospirillales;Rhodospirillaceae	0	0	1	28
Root;Bacteria;Proteobacteria;Alphaproteobacteria;Rickettsiales;Incertae sedis 4;Caedibacter	1	78	0	0
Root;Bacteria;Proteobacteria;Alphaproteobacteria;Rickettsiales;Incertae sedis 4;Odyssella	0	0	1	563
Root;Bacteria;Proteobacteria;Alphaproteobacteria;Sphingomonadales;Sphingomonadaceae	2	2571	0	0
Root;Bacteria;Proteobacteria;Betaproteobacteria	8	6783	3	65
Root;Bacteria;Proteobacteria;Betaproteobacteria;Burkholderiales	6	8931	0	0
Root;Bacteria;Proteobacteria;Betaproteobacteria;Burkholderiales;Burkholderiaceae	1	11	0	0
Root;Bacteria;Proteobacteria;Betaproteobacteria;Burkholderiales;Burkholderiaceae;Paucimonas	0	0	0	0
Root;Bacteria;Proteobacteria;Betaproteobacteria;Burkholderiales;Comamonadaceae	6	1759	0	0
Root;Bacteria;Proteobacteria;Betaproteobacteria;Burkholderiales;Comamonadaceae;Ramlibacter	0	0	1	44
Root;Bacteria;Proteobacteria;Betaproteobacteria;Burkholderiales;Oxalobacteraceae	0	0	2	493
Root;Bacteria;Proteobacteria;Betaproteobacteria;Burkholderiales;Oxalobacteraceae;Herbaspirillum	0	0	1	23
Root;Bacteria;Proteobacteria;Betaproteobacteria;Burkholderiales;Oxalobacteraceae;Herminiimonas	1	175	0	0
Root;Bacteria;Proteobacteria;Deltaproteobacteria	14	2942	6	1583
Root;Bacteria;Proteobacteria;Deltaproteobacteria;Bdellovibrionales;Bacteriovoracaceae	1	12	0	0
Root;Bacteria;Proteobacteria;Deltaproteobacteria;Bdellovibrionales;Bdellovibrionaceae;Bdellovibrio	2	222	0	0
Root;Bacteria;Proteobacteria;Deltaproteobacteria;Myxococcales	24	3009	4	434
Root;Bacteria;Proteobacteria;Deltaproteobacteria;Myxococcales;Cystobacterineae	4	415	5	217
Root;Bacteria;Proteobacteria;Deltaproteobacteria;Myxococcales;Cystobacterineae;Cystobacteraceae	1	147	0	0
Root;Bacteria;Proteobacteria;Deltaproteobacteria;Myxococcales;Cystobacterineae;Cystobacteraceae;Anaeromyxobacter	1	20	3	576
Root;Bacteria;Proteobacteria;Deltaproteobacteria;Myxococcales;Cystobacterineae;Myxococcaceae;Myxococcus	0	0	1	15
Root;Bacteria;Proteobacteria;Deltaproteobacteria;Myxococcales;Nannocystineae;Haliangiaceae;Haliangium	1	104	0	0
Root;Bacteria;Proteobacteria;Deltaproteobacteria;Myxococcales;Nannocystineae;N	1	38	0	0

annocystaceae				
Root;Bacteria;Proteobacteria;Deltaproteobacteria;Myxococcales;Nannocystineae;Nannocystaceae;Enhygromyxa	4	10021	0	0
Root;Bacteria;Proteobacteria;Deltaproteobacteria;Myxococcales;Sorangineae;Polyangiaceae	12	4711	4	1364
Root;Bacteria;Proteobacteria;Deltaproteobacteria;Myxococcales;Sorangineae;Polyangiaceae;Polyangium	6	590	7	609
Root;Bacteria;Proteobacteria;Gammaproteobacteria	29	110487	37	8384
Root;Bacteria;Proteobacteria;Gammaproteobacteria;Chromatiales	0	0	0	0
Root;Bacteria;Proteobacteria;Gammaproteobacteria;Legionellales;Coxiellaceae	1	70	0	0
Root;Bacteria;Proteobacteria;Gammaproteobacteria;Legionellales;Coxiellaceae;Aquicella	1	42	8	2737
Root;Bacteria;Proteobacteria;Gammaproteobacteria;Legionellales;Legionellaceae	1	265	3	499
Root;Bacteria;Proteobacteria;Gammaproteobacteria;Legionellales;Legionellaceae;Legionella	1	46	0	0
Root;Bacteria;Proteobacteria;Gammaproteobacteria;Pseudomonadales;Pseudomonadaceae	1	684	0	0
Root;Bacteria;Proteobacteria;Gammaproteobacteria;Pseudomonadales;Pseudomonadaceae;Azotobacter	1	53	0	0
Root;Bacteria;Proteobacteria;Gammaproteobacteria;Pseudomonadales;Pseudomonadaceae;Pseudomonas	1	51	0	0
Root;Bacteria;Proteobacteria;Gammaproteobacteria;Thiotrichales;Francisellaceae;Francisella	1	80	0	0
Root;Bacteria;Proteobacteria;Gammaproteobacteria;Xanthomonadales;Xanthomonadaceae	5	1332	1	2424
Root;Bacteria;Proteobacteria;Gammaproteobacteria;Xanthomonadales;Xanthomonadaceae;Hydrocarboniphaga	1	280	0	0
Root;Bacteria;Proteobacteria;Gammaproteobacteria;Xanthomonadales;Xanthomonadaceae;Lysobacter	3	2221	0	0
Root;Bacteria;Spirochaetes;Spirochaetes;Spirochaetales;Spirochaetaceae;Spirochaeta	0	0	1	374
Root;Bacteria;TM7;TM7_genera_incertae_sedis	1	9	0	0
Root;Bacteria;Verrucomicrobia;Verrucomicrobiae;Verrucomicrobiales	1	85	0	0
Root;Bacteria;Verrucomicrobia;Verrucomicrobiae;Verrucomicrobiales;Opitutaceae;Opitutus	0	0	1	203
Root;Bacteria;Verrucomicrobia;Verrucomicrobiae;Verrucomicrobiales;Subdivision 3;Subdivision 3_genera_incertae_sedis	12	1754	9	1931
Root;Bacteria;Verrucomicrobia;Verrucomicrobiae;Verrucomicrobiales;Verrucomicrobiaceae	2	609	0	0
Root;Bacteria;Verrucomicrobia;Verrucomicrobiae;Verrucomicrobiales;Verrucomicrobiaceae;Verrucomicrobiaceae_genera_incertae_sedis	1	10	0	0
Root;Bacteria;Verrucomicrobia;Verrucomicrobiae;Verrucomicrobiales;Xiphinematobacteriaceae;Xiphinematobacteriaceae_genera_incertae_sedis	0	0	2	98

Bibliography

<Kowalchuck 1997 AOB.pdf>.

Acinas, S.G., Marcelino, L.A., Klepac-Ceraj, V., and Polz, M.F. 2004. Divergence and Redundancy of 16S rRNA Sequences in Genomes with Multiple rrn Operons. J. Bacteriol. **186**: 2629-2635.

Ahmed, N., Gilbert, J.A., Field, D., Huang, Y., Edwards, R., Li, W., Gilna, P., and Joint, I. 2008. Detection of large numbers of novel sequences in the metatranscriptomes of complex marine microbial communities. PLOS ONE. **3**: e3042.

Amann, R., Fuchs, B.M., and Behrens, S. 2001. The identification of microorganisms by fluorescence *in situ* hybridisation. Curr. Opin. Biotech. **12**: 231-236.

Amann, R., and Ludwig, W. 2000. Ribosomal RNA-targeted nucleic acid probes for studies in microbial ecology. FEMS Microbiol. Rev. **24**: 555-565.

Amann, R.I., Ludwing, W., and Schleifer, K. 1995. Phylogenetic identification and *in situ* detection of individual microbial cells without cultivation. Microbiol. Rev. **59**: 143-169.

Anderson, C.R., Condron, L.M., Clough, T.J., Fiers, M., Stewart, A., Hill, R.A., and Sherlock, R.R. 2011. Biochar induced soil microbial community change: Implications for biogeochemical cycling of carbon, nitrogen and phosphorus. Pedobiologia. **54**: 309-320.

Andersson, A.F., Lindberg, M., Jakobsson, H., Bäckhed, F., Nyrén, P., and Engstrand, L. 2008. Comparative analysis of human gut microbiota by barcoded pyrosequencing. PLOS ONE. **3**: e2836.

Antal, M.J., and Gronli, M. 2003. The art, science, and technology of charcoal production. Ind. Eng. Chem. Res. **24**: 1619-1640.

Ashby, M.N., Rine, J., Mongodin, E.F., Nelson, K.E., and Dimster-Denk, D. 2007. Serial analysis of rRNA genes and the unexpected dominance of rare members of microbial communities. Appl. Environ. Microbiol. **73**: 4532-4542.

Atkinson, C.J., Fitzgerald, J.D., and Hipps, N.A. 2010. Potential mechanisms for achieving agricultural benefits from biochar application to temperate soils: a review. Plant Soil. **337**: 1-18.

Azargohar, R., Nanda, S., Kosinski, J.A., Dalai, A.K., and Sutarto, R. 2014. Effects of temperature on the physicochemical characteristics of fast pryrolysis bio-chars derived from canadian waste biomass. Fuel. **125**: 90-100.

Barns, S.M., Takala, S.L., and Kuske, C.R. 1999. Wide distribution and diversity of members of the bacterial kingdom Acidobacterium in the environment. Appl. Environ. Microbiol. **65**: 1731-1737.

Bartram, A.K., Lynch, M.D.J., Stearns, J.C., Moreno-Hagelsieb, G., and Neufeld, J.D. 2011. Generation of multi-million 16S rRNA gene libraries from complex microbial communities by assembling paired-end Illumina reads. Appl. Environ. Microbiol. **77**: 3846-3852.

Benson, D.A., Karsch-Mizrachi, I., Lipman, D.J., Ostell, J., and Wheeler, D.L. 2005. GenBank. Nucleic. Acids. Res. **33**: D34-38.

Blackwood, C.B., Hudleston, D., Zak, D.R., and Buyer, J.S. 2007. Interpreting ecological diversity indices applied to terminal restriction fragment length polymorphism data: insights from simulated microbial communities. Appl. Environ. Microbiol. **73**: 5276-5283.

Blow, N. 2008. Metagenomics: Exploring unseen communities. Nature. **453**: 687-690.

Bohannan, B.J.M., and Hughes, J. 2003. New approaches to analyzing microbial biodiversity data. Curr. Opin. Microbiol. **6**: 282-287.

Borneman, J., and Triplett, E.W. 1997. Molecular microbial diversity in soils from eastern Amazonia: evidence for unusual microorganisms and microbial population shifts associated with deforestation. Appl. Environ. Microbiol. **63**: 2647-2653.

Bougnom, B.P., Knapp, B.A., Elhottová, D., Koubová, A., Etoa, F.X., and Insam, H. 2010. Designer compost with biomass ashes for ameliorating acid tropical soils: Effects on the soil microbiota. Appl. Soil Ecol. **45**: 319-324.

Caporaso, J., Lauber, C., Walters, W., Berg-Lyons, D., Lozupone, C., Turnbaugh, P., Fierer, N., and Knight, R. 2010a. Global patterns of 16S rRNA diversity at a depth of millions of sequences per sample. Proc. Natl. Acad. Sci. U. S. A.

Caporaso, J.G., Bittinger, K., Bushman, F.D., DeSantis, T.Z., Anderson, G.L., and Knight, R. 2009. PyNAST: a flexible tool for aligning sequences to a template alignment. Bioinformatics. **26**: 266-267.

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

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

Caporaso, J.G., Lauber, C.L., Walters, W.A., Berg-Lyons, D., Lozupone, C.A., Turnbaugh, P.J., Fierer, N., and Knight, R. 2010c. Global patterns of 16S rRNA diversity at a depth of millions of sequences per sample. Proc. Natl. Acad. Sci. U.S.A. **108**: 4516-4522.

Chao, A. 1984. Non-parametric estimation of the number of classes in a population. Scand. J. Stat. **11**: 265-270.

Chao, A., and Yang, M.C.K. 1993. Stopping rules and estimation for recapture debugging with unequal failure rates. Biometrika. **80**: 193-291.

Chu, H., Fierer, N., Lauber, C.L., Caporaso, J.G., Knight, R., and Grogan, P. 2010. Soil bacterial diversity in the Arctic is not fundamentally different from that found in other biomes. Environ. Microbiol. **12**: 2998-3006.

Claesson, M.J., Wang, Q., O'Sullivan, O., Greene-Diniz, R., Cole, J.R., Ross, R.P., and O'Toole, P.W. 2010. Comparison of two next-generation sequencing technologies for resolving highly complex microbiota composition using tandem variable 16S rRNA gene regions. Nucleic Acids Res. **38**: e200-e200.

Cole, C.V., Duxbury, J., Freney, J., Heinemeyer, O., Minami, K., Mosier, A., Paustian, K., Rosenburg, N., Sampson, N., Sauerbect, D., and Zhao, Q. 1997. Global estimates of potential mitigation of greenhouse gas emissions by agriculture. Nutr. Cycling Agroecosyst. **49**: 221-228.

Cole, J.R., Wang, Q., Cardenas, E., Fish, J., Chai, B., Farris, R.J., Kulam-Syed-Mohideen, A.S., McGarrell, D.M., Marsh, T., Garrity, G.M., and Tiedje, J.M. 2009. The Ribosomal Database Project: improved alignments and new tools for rRNA analysis. Nucleic Acids Research **37**: D141-145.

Cole, J.R., Wang, Q., Fish, J.A., Chai, B., McGarrell, D.M., Sun, Y., Brown, C.T., Porras-Alfaro, A., Kuske, C.R., and Tiedje, J.M. 2014. Ribosomal Database Project: data and tools for high throughput rRNA analysis. Nucleic Acids Res. **42**: D633-642.

Curtis, T.P., Head, I.M., Lunn, M., Woodcock, S., Schloss, P.D., and Sloan, W.T. 2006. What is the extent of prokaryotic diversity? Philos. Trans. R. Soc. London, Ser. B **361**: 2023-2037.

Curtis, T.P., Sloan, W.T., and Scannell, J.W. 2002. Estimating prokaryotic diversity and its limits. Proc. Natl. Acad. Sci. U.S.A. **99**: 10494-10499.

Daims, H., Lücker, S., and Wagner, M. 2006. daime, a novel image analysis program for microbial ecology and biofilm research. Environ. Microbiol. **8**: 200-213.

del Giorgio, P.A., and Cole, J.J. 1998. Bacterial growth efficiency in natural aquatic systems. Annu. Rev. Ecol. Syst. **29**: 501-541.

DeSantis, T.Z., Jr., Hugenholtz, P., Keller, K., Brodie, E.L., Larsen, N., Piceno, Y.M., Phan, R., and Andersen, G.L. 2006. NAST: a multiple sequence alignment server for comparative analysis of 16S rRNA genes. Nucleic Acids Res. **34**: W394-399.

Dethlefsen, L., Huse, S., Sogin, M.L., and Relman, D.A. 2008. The pervasive effects of an antibiotic on the human gut microbiota, as revealed by deep 16S rRNA sequencing. PLoS Biology. **6**: e280.

Doolittle, W.F. 1999. Phylogenetic classification and the universal tree. Science. **284**: 2124-2128.

Doroghazi, J.R., and Buckley, D.H. 2008. Evidence from GC-TRFLP that bacterial communities in soil are lognormally distributed. PLOS ONE. **3**: e2910.

Dufrene, M., and Legendre, P. 1997. Species assemblages and indicator species: the need for a flexable asymmetrical approach. Ecol. Monogr. **67**: 345-366.

Dunbar, J., Ticknor, L.O., and Kuske, C.R. 2000. Assessment of microbial diversity in four southwestern United States soils by 16S rRNA gene terminal restriction fragment analysis. Appl. Environ. Microbiol. **66**: 2943-2950.

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

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

Eilers, K.G., Debenport, S., Anderson, S.A., and Fierer, N. 2012. Digging deeper to find unique microbial communities: The strong effect of depth on the structure of bacterial and archaeal communities in soil. Soil Biol. Biochem. **50**: 58-65.

Favet, J., Lapanje, A., Giongo, A., Kennedy, S., Aung, Y.Y., Cattaneo, A., Davis-Richardson, A.G., Brown, C.T., Kort, R., Brumsack, H.J., Schnetger, B., Chappell, A., Kroijenga, J., Beck, A., Schwibbert, K., Mohamed, A.H., Kirchner, T., de Quadros, P.D., Triplett, E.W., Broughton, W.J., and Gorbushina, A.A. 2013. Microbial hitchhikers on intercontinental dust: catching a lift in Chad. ISME J **7**: 850-867.

Fierer, N. 2007. Tilting at windmills: a response to a recent critique of terminal restriction fragment length polymorphism data. Appl. Environ. Microbiol. **73**: 8041; author reply 8041-8042.

Fierer, N., and Jackson, R.B. 2006. The diversity and biogeography of soil bacterial communities. Proc. Natl. Acad. Sci. U.S.A. **103**: 626-631.

Fox, G.E., Stackebrandt, E., Hespell, R.B., Gibson, J., Maniloff, J., Dyer, T.A., Wolfe, R.S., Balch, W.E., Tanner, R.S., Magrum, L.J., Zablen, L.B., Blakemore, R., Gupta, R., Bonen, L., Lewis, B.J., Stahl, D.A., Luehrsen, K.R., Chen, K.N., and Woese, C.R. 1980. The phylogeny of prokaryotes. Science. **209**: 457-463.

Frank, D.N. 2009. BARCRAWL and BARTAB: software tools for the design and implementation of barcoded primers for highly multiplexed DNA sequencing. BMC Bioinf. **10**: 362.

Franklin, R.B., and Mills, A.L. 2006. Structural and functional responses of a sewage microbial community to dilution-induced reductions in diversity. Microb. Ecol. **52**: 280-288.

Fulthorpe, R.R., Roesch, L.F., Riva, A., and Triplett, E.W. 2008. Distantly sampled soils carry few species in common. ISME J. **2**: 901-910.

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

Glaser, B., Haumaier, L., Guggenberger, G., and Zech, W. 2001. The 'Terra Preta' phenomenon: a model for sustainable agriculture in the tropics. Naturwissenschaften. **88**: 37-41.

Gloor, G.B., Hummelen, R., Macklaim, J.M., Dickson, R.J., Fernandes, A.D., MacPhee, R., and Reid, G. 2010. Microbiome profiling by Illumina sequencing of combinatorial sequence-tagged PCR products. PLOS ONE. **5**: e15406.

Goebel, B.M., and Stackebrandt, E. 1994. Cultural and phylogenetic analysis of mixed microbial populations found in natural and commercial bioleaching environments. Appl. Environ. Microbiol. **60**: 1614-1621.

Gonzalez, J.M., Zimmermann, J., and Saiz-Jimenez, C. 2005. Evaluating putative chimeric sequences from PCR-amplified products. Bioinformatics. **21**: 333-337.

Good, I.J. 1953. The population frequencies of species and the estimation of population parameters. Biometrika. **40**: 237-264.

Green, S.J., Leigh, M.B., and Neufeld, J.D. 2010. Denaturing gradient gel electrophoresis (DGGE) for microbial community analysis. *In* Microbiology of Hydrocarbon and Lipid Microbiology. *Edited by* K.N. Timmis. Springer-Verlag Berlin Heidelberg, Berlin. pp. 4137-4158.

Green, S.J., Prakash, O., Jasrotia, P., Overholt, W.A., Cardenas, E., Hubbard, D., Tiedje, J.M., Watson, D.B., Schadt, C.W., Brooks, S.C., and Kostka, J.E. 2012. Denitrifying bacteria from the genus Rhodanobacter dominate bacterial communities in the highly contaminated subsurface of a nuclear legacy waste site. Appl. Environ. Microbiol. **78**: 1039-1047.

Grossman, J.M., O'Neill, B.E., Tsai, S.M., Liang, B., Neves, E., Lehmann, J., and Thies, J.E. 2010. Amazonian anthrosols support similar microbial communities that differ distinctly from those extant in adjacent, unmodified soils of the same mineralogy. Microb. Ecol. **60**: 192-205.

Hamady, M., Walker, J.J., Harris, J.K., Gold, N.J., and Knight, R. 2008. Error-correcting barcoded primers for pyrosequencing hundreds of samples in multiplex. Nat. Methods. **5**: 235-237.

Hill, J.E., Seipp, R.P., Betts, M., Hawkins, L., Van Kessel, A.G., Crosby, W.L., and Hemmingsen, S.M. 2002. Extensive profiling of a complex microbial community by high-throughput sequencing. Appl. Environ. Microbiol. **68**: 3055-3066.

Hua, L., Lu, Z., Ma, H., and Jin, S. 2014. Effect of biochar on carbon dioxide release, organic carbon accumulation, and aggregation of soil. Environ. Prog. Sustainable Energy. **33**: 941-946.

Huber, J.A., Welch, D.B.M., Morrison, H.G., Huse, S.M., Neal, P.R., Butterfield, D.A., and Sogin, M.L. 2007. Microbial population structures in the deep marine biosphere. Science. **318**: 97-100.

Hugerth, L.W., Muller, E.E., Hu, Y.O., Lebrun, L.A., Roume, H., Lundin, D., Wilmes, P., and Andersson, A.F. 2014. Systematic design of 18S rRNA gene primers for determining eukaryotic diversity in microbial consortia. PLoS ONE. **9**: e95567.

Hughes, J.B., Hellmann, J.J., Ricketts, T.H., and Bohannan, B.J.M. 2001. Counting the uncountable: statistical approaches to estimating microbial diversity. Appl. Environ. Microbiol. **67**: 4399-4406.

Hummelen, R., Fernandes, A.D., Macklaim, J.M., Dickson, R.J., Changalucha, J., Gloor, G.B., and Reid, G. 2010. Deep sequencing of the vaginal microbiota of women with HIV. PLOS ONE. **5**: e12078.

Huse, S.M., Dethlefsen, L., Huber, J.A., Welch, D.M., Relman, D.A., and Sogin, M.L. 2008. Exploring microbial diversity and taxonomy using SSU rRNA hypervariable tag sequencing. PLoS Genet. **4**: e1000255.

Huse, S.M., Welch, D.M., Morrison, H.G., and Sogin, M.L. 2010. Ironing out the wrinkles in the rare biosphere through improved OTU clustering. Environ. Microbiol. **12**: 1889-1898.

Janvier, M., Regnault, B., and Grimont, P. 2003. Development and use of fluorescent 16S rRNA-targeted probes for the specific detection of *Methylophaga* species by *in situ* hybridization in marine sediments. Res. Microbiol. **154**: 483-490.

Jiang, X., Langille, M.G., Neches, R.Y., Elliot, M., Levin, S.A., Eisen, J.A., Weitz, J.S., and Dushoff, J. 2012. Functional biogeography of ocean microbes revealed through non-negative matrix factorization. PLoS ONE. **7**: e43866.

Jiang, X., Weitz, J.S., and Dushoff, J. 2011. A non-negative matrix factorization framework for identifying modular patterns in metagenomic profile data. J. Math Biol.

Jolley, K.A., Bliss, C.M., Bennett, J.S., Bratcher, H.B., Brehony, C., Colles, F.M., Wimalarathna, H., Harrison, O.B., Sheppard, S.K., Cody, A.J., and Maiden, M.C. 2012. Ribosomal multilocus sequence typing: universal characterization of bacteria from domain to strain. Microbiology **158**: 1005-1015.

Jones, R.T., Robeson, M.S., Lauber, C.L., Hamady, M., Knight, R., and Fierer, N. 2009. A comprehensive survey of soil acidobacterial diversity using pyrosequencing and clone library analyses. ISME J. **3**: 442-453.

Kennedy, K., Hall, M.W., Lynch, M.D.J., Moreno-Hagelsieb, G., and Neufeld, J.D. 2014. Evaluating bias of Illumina-based bacterial 16S rRNA gene profiles. Appl. Environ. Microbiol. **80**: 5717-5722.

Kirk, J.L., Klironomos, J.N., Lee, H., and Trevors, J.T. 2005. The effects of perennial ryegrass and alfalfa on microbial abundance and diversity in petroleum contaminated soil. Environ. Pollut. **133**: 455-465.

Klappenbach, J.A., Dunbar, J.M., and Schmidt, T.M. 2000. rRNA operon copy number reflects ecological strategies of bacteria. Appl. Env. Microbiol. **66**: 1328-1333.

Kuczynski, J., Costello, E.K., Nemergut, D.R., Zaneveld, J., Lauber, C.L., Knights, D., Koren, O., Fierer, N., Kelley, S.T., Ley, R.E., Gordon, J.I., and Knight, R. 2010. Direct sequencing of the human microbiome readily reveals community differences. Genome Biol. **11**: 210.

Kunin, V., Engelbrektson, A., Ochman, H., and Hugenholtz, P. 2010. Wrinkles in the rare biosphere: pyrosequencing errors can lead to artificial inflation of diversity estimates. Environ. Microbiol. **12**: 118-123.

Kysela, D.T., Palacios, C., and Sogin, M.L. 2005. Serial analysis of V6 ribosomal sequence tags (SARST-V6): a method for efficient, high-throughput analysis of microbial community composition. Environ. Microbiol. **7**: 356-364.

Lal, R. 2008. Carbon sequestration. Phil. Trans. R. Soc. B. **363**: 815-830.

Lane, D.J. 1991. 16S/23S rRNA sequencing. *In* Nucleic Acid Tech. Bact. Syst. *Edited by* E. Stackebrandt and M. Goodfellow. John Wiley & Sons, Inc., Chichester, UK. pp. 115-175.

Langenheder, S., and Prosser, J.I. 2008. Resource availability influences the diversity of a functional group of heterotrophic soil bacteria. Environ. Microbiol. **10**: 2245-2256.

Lauber, C.L., Hamady, M., Knight, R., and Fierer, N. 2009. Pyrosequencing-based assessment of soil pH as a predictor of soil bacterial community structure at the continental scale. Appl. Environ. Microbiol. **75**: 5111-5120.

Lazarevic, V., Whiteson, K., Huse, S., Hernandez, D., Farinelli, L., Osteras, M., Schrenzel, J., and Francois, P. 2009. Metagenomic study of the oral microbiota by Illumina high-throughput sequencing. J. Microbiol. Methods. **79**: 266-271.

Lehmann, J., and Joseph, S. 2009. Biochar for environmental management: science and technology. Earthscan, London, UK. .

Lehmann, J., Rillig, M.C., Thies, J., Masiello, C.A., Hockaday, W.C., and Crowley, D. 2011. Biochar effects on soil biota – A review. Soil Biol. Biochem. **43**: 1812-1836.

Lehtovirta, L.E., Prosser, J.I., and Nicol, G.W. 2009. Soil pH regulates the abundance and diversity of Group 1.1c Crenarchaeota. FEMS Microbiol. Ecol. **70**: 367-376.

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

Liang, B.L., Lehmann, J., Solomon, D., Kinyangi, J., Grossman, J., O'Neill, B., Skjemstad, J.O., Thies, J., Luizao, F.J., Peterson, J., and Neves, E.G. 2006. Black carbon increases cation exchange capacity in soils. Soil. Sci. Soc. Am. J. **70**: 1719-1730.

Liu, L., Shen, G., Sun, M., Cao, X., Shang, G., and Chen, P. 2014. Effect of biochar on nitrous oxide emission and its potential mechanisms. J. Air Waste Manage. Assoc. **64**: 894-902.

Liu, W., Marsh, T.L., Cheng, H., and Forney, L.J. 1997. Characterization of microbial diversity by determining terminal restriction fragment length polymorphisms of genes encoding 16S rRNA. Appl. Env. Microbiol. **63**: 4516-4522.

Lozupone, C., Hamady, M., and Knight, R. 2006. UniFrac - An online tool for comparing microbisl community diversity in a phylogenetic context. BMC Bioinf. **7**: 371.

Lozupone, C., Lladser, M.E., Knights, D., Stombaugh, J., and Knight, R. 2010. UniFrac: an effective distance metric for microbial community comparison. ISME J. **5**: 169-172.

Lozupone, C.A., and Knight, R. 2007. Global patterns in bacterial diversity. Proc. Natl. Acad. Sci. U.S.A. **104**: 11436.

Lu, S., Gischkat, S., Reiche, M., Akob, D.M., Hallberg, K.B., and Kusel, K. 2010. Ecophysiology of Fe-cycling bacteria in acidic sediments. Appl. Environ. Microbiol. **76**: 8174-8183.

Luo, Q., Hiessl, S., and Steinbuchel, A. 2014. Functional diversity of Nocardia in metabolism. Environ Microbiol **16**: 29-48.

Lynch, M.D., Bartram, A.K., and Neufeld, J.D. 2012. Targeted recovery of novel phylogenetic diversity from next-generation sequence data. ISME J. **6**: 2067-2077.

Lynch, M.D.J., Masella, A.P., Hall, M.W., Bartram, A.K., and Neufeld, J.D. 2013. AXIOME: automated exploration of microbial diversity. GigaScience. **2**: 1-5.

Mao, J.D., Johnson, R.L., Lehmann, J., Olk, D.C., Neves, E.G., Thompson, M.L., and Schmidt-Rohr, K. 2012. Abundant and stable char residues in soils: implications for soil fertility and carbon sequestration. Environ. Sci. Technol. **46**: 9571-9576.

Masella, A.P., Bartram, A.K., Truszkowski, J.M., Brown, D.G., and Neufeld, J.D. 2012. PANDAseq: paired-end assembler for illumina sequences. BMC Bioinf. **13**: 31.

Maughan, H., Wang, P.W., Diaz Caballero, J., Fung, P., Gong, Y., Donaldson, S.L., Yuan, L., Keshavjee, S., Zhang, Y., Yau, Y.C., Waters, V.J., Tullis, D.E., Hwang, D.M., and Guttman, D.S. 2012. Analysis of the cystic fibrosis lung microbiota via serial Illumina sequencing of bacterial 16S rRNA hypervariable regions. PLOS ONE. **7**: e45791.

McMahon, K.D., Martin, H.G., and Hugenholtz, P. 2007. Integrating ecology into biotechnology. Curr. Opin. Biotechnol. **18**: 287-292.

Mohan, D., Pittman, C.U., and Steele, P.H. 2006. Pyrolysis of wood/biomass for bio-oil: a critical review. Energy Fuels. **20**: 848-889.

Mühling, M., Woolven-Allen, J., Murrell, J.C., and Joint, I. 2008. Improved group-specific PCR primers for denaturing gradient gel electrophoresis analysis of the genetic diversity of complex microbial communities. ISME J. **2**: 379-392.

Muyzer, G., De Waal, E.C., and Uitterlinden, A.G. 1993. Profiling of complex microbial populations by denaturing gradient gel electrophoresis analysis of polymerase chain reaction-amplified genes coding for 16S rRNA. Appl. Environ. Microbiol. **59**: 695-700.

Nawrocki, E.P., Kolbe, D.L., and Eddy, S.R. 2009. Infernal 1.0: inference of RNA alignments. Bioinformatics. **25**: 1335-1337.

Neefs, J.M., Van de Peer, Y., De Rijk, P., Chapelle, S., and De Wachter, R. 1993. Compilation of small ribosomal subunit RNA structures. Nucl. Acids Res. **21**: 3025-3049.

Neufeld, J.D., Li, J., and Mohn, W.W. 2008. Scratching the surface of the rare biosphere with ribosomal sequence tag primers. FEMS Microbiol. Lett. **283**: 146-153.

Neufeld, J.D., and Mohn, W.W. 2005a. Assessment of microbial phylogenetic diversity based on environmental nucleic acids. *In* Molecular Identification, Systematics, and Population Structure of Prokaryotes. *Edited by* E. Stackebrandt. Springer-Verlag, Heidelberg.

Neufeld, J.D., and Mohn, W.W. 2005b. Unexpectedly high bacterial diversity in arctic tundra relative to boreal forest soils revealed with serial analysis of ribosomal sequence tags (SARST). Appl. Environ. Microbiol. **71**: 5710-5718.

Neufeld, J.D., and Mohn, W.W. 2006. Assessment of microbial phylogenetic diversity based on environmental nucleic acids. In Molecular Identification, Systematics, and Population Structure of Prokaryotes. Heidelberg: Springer-Verlag.

Neufeld, J.D., Yu, Z., Lam, W., and Mohn, W.W. 2004. Serial analysis of ribosomal sequence tags (SARST): a high-throughput method for profiling complex microbial communities. Environ. Microbiol. **6**: 131-144.

Nicol, G.W., Leininger, S., Schleper, C., and Prosser, J.I. 2008. The influence of soil pH on the diversity, abundance and transcriptional activity of ammonia oxidizing archaea and bacteria. Environ. Microbiol. **10**: 2966-2978.

Nielsen, S., Minchin, T., Kimber, S.W.L., Van Zwieten, L., Gilbert, J., Munroe, P., Joseph, S., and Thomas, T. 2014. Comparative analysis of the microbial communities in agricultural soil amended with enhanced biochars or traditional fertilisers. Agric. Ecosyst. Environ. **191**: 73-82.

Novak, J.M., Lima, I., Xing, B., Gaskin, J.W., Steiner, C., Das, K.C., Ahmedna, M., Djaafar, R., Watts, D.W., Busscher, W.J., and Schomberg, H. 2009. Characterization of designer biochar produced at different temperatures and their effects on a loamy sand. Ann. Environ. Sci. **3**: 195-206.

Nüsslein, K., and Tiedje, J.M. 1998. Characterization of the dominant and rare members of a young Hawaiian soil bacterial community with small-subunit ribosomal DNA amplified from

DNA fractionated on the basis of its guanine and cytosine composition. Appl. Environ. Microbiol. **64**: 1283-1289.

Nyyssonen, M., Hultman, J., Ahonen, L., Kukkonen, I., Paulin, L., Laine, P., Itavaara, M., and Auvinen, P. 2014. Taxonomically and functionally diverse microbial communities in deep crystalline rocks of the Fennoscandian shield. ISME J. **8**: 126-138.

O'Neill, B., Grossman, J., Tsai, M.T., Gomes, J.E., Lehmann, J., Peterson, J., Neves, E., and Thies, J.E. 2009. Bacterial community composition in Brazilian Anthrosols and adjacent soils characterized using culturing and molecular identification. Microb. Ecol. **58**: 23-35.

Olsen, G.J., Lane, D.J., Giovannoni, S.J., Pace, N.R., and Stahl, D.A. 1986. Microbial ecology and evolution: a ribosomal RNA approach. Annu. Rev. Microbiol. **40**: 337-365.

Ong, S.H., Kukkillaya, V.U., Wilm, A., Lay, C., Ho, E.X., Low, L., Hibberd, M.L., and Nagarajan, N. 2013. Species identification and profiling of complex microbial communities using shotgun Illumina sequencing of 16S rRNA amplicon sequences. PLOS ONE. **8**: e60811.

Paul, E.A. 2007. Soil microbiology, ecology, and biochemistry. 3 ed. Academic Press, Amsterdam; Boston.

Pedros-Alio, C. 2007. Dipping into the rare biosphere. Science. **315**: 192-193.

Postma, J., Schilder, M.T., and van Hoof, R.A. 2011. Indigenous populations of three closely related *Lysobacter* spp. in agricultural soils using real-time PCR. Microb. Ecol. **62**: 948-958.

Prosser, J.I., Bohannan, B.J.M., Curtis, T.P., Ellis, R.J., Firestone, M.K., Freckleton, R.P., Green, J.L., Green, L.E., Killham, K., Lennon, J.J., Osborn, A.M., Solan, M., van der Gast, C.J., and Young, J.P.W. 2007. The role of ecological theory in microbial ecology. Nat. Rev. Microbiol. **5**: 384-392.

Rappe, M.S., and Giovannoni, S.J. 2003. The uncultured microbial majority. Annu. Rev. Microbiol. **57**: 369-394.

Rawat, S.R., Mannisto, M.K., Bromberg, Y., and Haggblom, M.M. 2012. Comparative genomic and physiological analysis provides insights into the role of Acidobacteria in organic carbon utilization in Arctic tundra soils. FEMS Microbiol. Ecol. **82**: 341-355.

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

Rousk, J., Baath, E., Brookes, P.C., Lauber, C.L., Lozupone, C., Caporaso, J.G., Knight, R., and Fierer, N. 2010. Soil bacterial and fungal communities across a pH gradient in an arable soil. ISME J. **4**: 1340-1351.

Rousk, J., Brookes, P.C., and Baath, E. 2009. Contrasting soil pH effects on fungal and bacterial growth suggest functional redundancy in carbon mineralization. Appl. Environ. Microbiol. **75**: 1589-1596.

Russo, S.E., Legge, R., Weber, K.A., Brodie, E.L., Goldfarb, K.C., Benson, A.K., and Tan, S. 2012. Bacterial community structure of contrasting soils underlying Bornean rain forests: Inferences from microarray and next-generation sequencing methods. Soil Biol. Biochem. **55**: 48-59.

Rutherford, D.W., Wershaw, R.L., Rostad, C.E., and Kelly, C.N. 2012. Effect of formation conditions on biochars: Compositional and structural properties of cellulose, lignin, and pine biochars. Biomass Bioenergy. **46**: 693-701.

Saiki, R.K., Scharf, S., Faloona, F., Mullis, K.B., Horn, G.T., Erlich, H.a., and Arnheim, N. 1985. Enzymatic amplification of beta-globin genomic sequences and restriction site analysis for diagnosis of sickle cell anemia. Science. **230**: 1350-1354.

Sbpner, A., Mu, X.J., Greenbaum, D., Auerbach, R.K., and Gerstein, M.B. 2011. The real cost of sequencing: higher than you think! Genome Biol. **12**: 125.

Schloss, P., and Handelsman, J. 2003. Biotechnological prospects from metagenomics. Curr. Opin. Biotechnol. **14**: 303-310.

Schloss, P.D., and Handelsman, J. 2004. Status of the microbial census. Microbiol. Mol. Biol. Rev. **68**: 686-691.

Shendure, J., and Ji, H. 2008. Next-generation DNA sequencing. Nat. Biotechnol. **26**: 1135-1145.

Shu, Q., and Jiao, N. 2013. Developing a novel approach of *rpoB* gene as a powerful biomarker for the environmental microbial diversity. Geomicrobiol. J. **30**: 108-119.

Singh, B.K., and Macdonald, C.A. 2010. Drug discovery from uncultivable microorganisms. Drug Discovery Today. **15**: 792-799.

Smith, P., Haberl, H., Popp, A., Erb, K.H., Lauk, C., Harper, R., Tubiello, F.N., de Siqueira Pinto, A., Jafari, M., Sohi, S., Masera, O., Bottcher, H., Berndes, G., Bustamante, M., Ahammad, H., Clark, H., Dong, H., Elsiddig, E.A., Mbow, C., Ravindranath, N.H., Rice, C.W., Robledo Abad, C., Romanovskaya, A., Sperling, F., Herrero, M., House, J.I., and Rose, S. 2013. How much land-based greenhouse gas mitigation can be achieved without compromising food security and environmental goals? Global Change Biol. **19**: 2285-2302.

Sogin, M.L., Morrison, H.G., Huber, J.A., Mark Welch, D., Huse, S.M., Neal, P.R., Arrieta, J.M., and Herndl, G.J. 2006. Microbial diversity in the deep sea and the underexplored "rare biosphere". Proc. Natl. Acad. Sci. U.S.A. **103**: 12115-12120.

Solomon, S., Plattner, G.K., Knutti, R., and Friedlingstein, P. 2009. Irreversible climate change due to carbon dioxide emissions. Proc. Natl. Acad. Sci. U.S.A. **106**: 1704-1709.

Stach, J.E.M., Maldonado, L.A., Ward, A.C., Goodfellow, M., and Bull, A.T. 2003. New primers for the class Actinobacteria: application to marine and terrestrial environments. Environ. Microbiol. **5**: 828-841.

Steinbeiss, S., Gleixner, G., and Antonietti, M. 2009. Effect of biochar amendment on soil carbon balance and soil microbial activity. Soil Biol. Biochem. **41**: 1301-1310.

Stevenson, B.S., Eichorst, S.A., Wertz, J.T., Schmidt, T.M., and Breznak, J.A. 2004. New strategies for cultivation and detection of previously uncultured microbes. Appl. Environ. Microbiol. **70**: 4748-4755.

Taketani, R.G., Lima, A.B., da Conceicao Jesus, E., Teixeira, W.G., Tiedje, J.M., and Tsai, S.M. 2013. Bacterial community composition of anthropogenic biochar and Amazonian anthrosols assessed by 16S rRNA gene 454 pyrosequencing. Antonie van Leeuwenhoek **104**: 233-242.

Torsvik, V., Daae, F.L., Sandaa, R.A., and Ovreas, L. 1998. Novel techniques for analysing microbial diversity in natural and perturbed environments. J. Biotechnol. **64**: 53-62.

Tyler, H.L., Khalid, S., Jackson, C.R., and Moore, M.T. 2013. Determining potential for microbial atrazine degradation in agricultural drainage ditches. J. Environ. Qual. **42**: 828-834.

Vetriani, C., Voordeckers, J.W., Crespo-Medina, M., O'Brien, C.E., Giovannelli, D., and Lutz, R.A. 2014. Deep-sea hydrothermal vent Epsilonproteobacteria encode a conserved and widespread nitrate reduction pathway (Nap). ISME J. **8**: 1510-1515\1521.

Vogel, T.M. 2009. TerraGenome: a consortium for the sequencing of a soil metagenome. Nat. Rev. Micro. **7**: 252.

Voget, S., Steele, H.L., and Streit, W.R. 2006. Characterization of a metagenome-derived halotolerant cellulase. J. Biotechnol. **126**: 26-36.

Waluikar, S., Dhotre, D.P., Marathe, N.P., Lawate, P.S., Bharadwaj, R.S., and Shouche, Y.S. 2014. Characterization of bacterial community shift in human ulcerative colitis patients revealed by Illumina based 16S rRNA gene amplicon sequncing. Gut. Pathog. **6**: 1-11.

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

Whittaker, R.H. 1972. Evolution and measurement of species diversity. Taxon. **21**: 213-251.

Willis, K.J., and Whittaker, R.J. 2002. Ecology. Species diversity - scale matters. Science. **295**: 1245-1248.

Woese, C.R. 1987. Bacterial evolution. Microbiol. Rev. **51**: 221-271.

Woolf, D., Amonette, J.E., Street-Perrott, F.A., Lehmann, J., and Joseph, S. 2010. Sustainable biochar to mitigate global climate change. Nat Commun **1**: 56.

Yu, Z., Yu, M., and Morrison, M. 2006. Improved serial analysis of V1 ribosomal sequence tags (SARST-V1) provides a rapid, comprehensive, sequence-based characterization of bacterial diversity and community composition. Environ. Microbiol. **8**: 603-611.

Zhou, H.-W., Li, D.-F., Tam, N.F.-Y., Jiang, X.-T., Zhang, H., Sheng, H.-F., Qin, J., Liu, X., and Zou, F. 2010a. BIPES, a cost-effective high-throughput method for assessing microbial diversity. ISME J. **5**: 741-749.

Zhou, H.W., Li, D.F., Tam, N.F., Jiang, X.T., Zhang, H., Sheng, H.F., Qin, J., Liu, X., and Zou, F. 2010b. BIPES, a cost-effective high-throughput method for assessing microbial diversity. ISME J. **5**: 741-749.

Zhou, J., Xia, B., Treves, D.S., Wu, L.Y., Marsh, T.L., O'Neill, R.V., Palumbo, A.V., and Tiedje, J.M. 2002. Spatial and resource factors influencing high microbial diversity in soil. Appl. Environ. Microbiol. **68**: 326-334.