MOLECULAR CHARACTERIZATION OF WETLAND SOIL BACTERIAL COMMUNITIES IN CONSTRUCTED MESOCOSMS

THESIS

Elisabeth M. Leon, Captain, USAF

AFIT/GES/ENV/08-M04

DEPARTMENT OF THE AIR FORCE
AIR UNIVERSITY

AIR FORCE INSTITUTE OF TECHNOLOGY

Wright-Patterson Air Force Base, Ohio

The views expressed in this thesis are those of the author and do not reflect the official policy or position of the United States Air Force, Department of Defense, or the U.S. Government.

AFIT/GES/ENV/08-M04

MOLECULAR CHARACTERIZATION OF WETLAND SOIL BACTERIAL COMMUNITIES IN CONSTRUCTED MESOCOSMS

THESIS

Presented to the Faculty

Department of Systems and Engineering Management

Graduate School of Engineering and Management

Air Force Institute of Technology

Air University

Air Education and Training Command

In Partial Fulfillment of the Requirements for the

Degree of Master of Science in Environmental Engineering and Science

Elisabeth M. Leon, BS

Captain, USAF

March 2008

MOLECULAR CHARACTERIZATION OF WETLAND SOIL BACTERIAL

COMMUNITIES IN CONSTRUCTED MESOCOSMS

Elisabeth M. Leon, BS
Captain, USAF

Approved:

__//SIGNED//_________________________ ________
Charles A. Bleckmann (Chairman) Date

__//SIGNED//_________________________ ________
Stephanie A. Smith (Member) Date

__//SIGNED//_________________________ ________
Sonia E. Leach, Maj, USAF (Member) Date

AFIT/GES/ENV/08-M04

Abstract

This research characterized the effects of three species of wetland plant on the composition and diversity of the rhizosphere bacterial communities they supported. Diversity and community composition were addressed in relation to three factors: plant presence, plant species, and soil depth; these factors helped identify the diversity and composition of subsurface flow wetlands and its remediation potential. The largest sample of 16S rRNA DNA sequences ever collected to date was described here, and enabled us to make comparisons of the effects of the presence or absence of plants, plant species, and plant rhizosphere depth on microbial diversity and community composition, using newly developed software packages. It was determined that plant rhizosphere supported a more diverse microbial community than plant-free soils. Also there was evidence that *Eleocharis erythropoda* was significantly more diverse than the *Carex comosa* microbial community, but not significantly in comparison to the *Scirpus atrovirens* community. Samples were taken from a top, middle, and bottom layer. While there did not appear to be an effect of diversity due to depth, one of the three plant species did support a less diverse community at its middle depth than the other two plants. This finding was consistent with a previous wetland study, and was significant because wetlands planted with this species can promote a less diverse microbial community. The compositions based on phyla classifications by RDP of the communities, however, were not significant for any of the comparisons.

Acknowledgments

I would first and foremost like to thank my wonderful family. They have supported me unconditionally throughout my life and have been my motivation through this entire research experience. My mother has always been my sounding board and has showered me with unconditional love. My father is my role model in my career, and his guidance and advice are ever so important to me. Secondly I would like to express my sincere appreciation to my faculty advisor for his guidance and support throughout the course of this thesis effort. The insight and experience was certainly appreciated. I would, also, like to thank my wonderful thesis advisor; she has helped teach me the world of biology and her time and patience were so much appreciated. Also I would like to send a special thank you to several students that helped with laboratory efforts. A big thank you goes to the members of my laboratory. I also would like to thank my fellow AFIT students. My experience here at AFIT has been a wonderful time and I am truly blessed to have shared it with an extraordinary group of individuals. I look forward to serving with you all again real soon. Last but not least I would like to thank my God for blessing me with countless opportunities in life and the wonderful people who you have brought in my path.

Elisabeth M. Leon

Table of Contents

Page

Abstract ... iv

Acknowledgments ... v

Table of Contents ... vi

List of Figures ... viii

List of Tables ... ix

Chapter I: Introduction ... 12

Research Objectives ... 15

Chapter II: Literature Review ... 17

Treatment Wetlands and Microbial/Plant Interactions ... 17

Soil Microbial Diversity and Diversity Statistics ... 20

16S rRNA Gene Analysis Method ... 21

RDP and Alignment ... 25

Comparative Analysis and Software ... 26

Error ... 30

Statistical Analysis ... 32

Chapter III: Methodology ... 33

Experimental Overview ... 33

Nomenclature ... 36

Laboratory procedures ... 37

Quality Check for Laboratory Procedures ... 38

*Eco*R1 Restriction Enzyme Digestion and Gel Electrophoresis ... 39

Sequencing and Trimming ... 41

Analysis ... 51

Chapter IV: Results and Analysis ... 54

Overview ... 54

Research Objective 1 ... 61

Page

Research Objective 2 67

Research Objective 3 74

Chapter V: Conclusions and Recommendations..82

Overview 82

Research Question 1 83

Research Question 2 85

Research Question 3 87

Limitations of research 89

Significance of Research 90

Further Research...90

Summary...91

Appendix A: PCR Protocol Using HotStarTaq Master Mix..93

Appendix B: Cloning Month Legend 95

Appendix C: Mo Bio PowerSoil™ DNA Isolation Kit Extraction Protocol 96

Appendix D: StrataClone™ PCR Cloning Kit 100

Appendix E: Plamsid Prep Protocol 104

Appendix F: Restriction Digest Protocol...107

Appendix G: Gels 110

Appendix H: Richness Estimator and Rarefaction Curves 114

Bibliography 132

Vita..138

List of Figures

Table Page

Figure 1: Mesocosm Design 34
Figure 2: Gel from Ap53.L1 40
Figure 3: Editing Step 1 43
Figure 4: Editing Step 2 45
Figure 5: Editing Step 3 47
Figure 6: Editing Step 4 48
Figure 7: Edited and Trimmed Sequences 48
Figure 8: Schematic of Sequence Analysis 50
Figure 9: Phyla Classification for all Data using RDP Classifier 57
Figure 10: All Richness Estimates and Rarefaction Curves 59
Figure 11: Good's Coverage 60
Figure 12: Evenness 61
Figure 13: Phyla Classification Control vs Planted sequences using RDP Classifier. 62
Figure 14: Control and Planted Data Richness Estimates and Rarefaction Curves 65
Figure 15: Phylum Level Rarefaction Curve for Control and Planted Data 66
Figure 16: Phyla Classification for Plant Species using RDP Classifier 68
Figure 17: Plant Species Data Richness Estimates and Rarefaction Curves 71
Figure 18: Phylum Level Plant Species Data Rarefaction Curve 73
Figure 19: Phyla Classification for all Depth sequences using RDP Classifier 75
Figure 20: Depth Data Richness Estimates and Rarefaction Curves 78
Figure 21: Community Phylum Level Depth Rarefaction Curves 80

List of Tables

Table Page

Table 1: Mesocosm Plantings (Bishop 2006) 35
Table 2: Sequence Breakout 55
Table 3: Frequency Distribution of OTUs at 97% Similarity 55
Table 4: Frequency Distribution of OTUs at 80% Similarity 56
Table 5: Phyla Classification Percentages (Control vs. Planted) 63
Table 6: Phyla Classification Percentages for Plant Species 69
Table 7: Phyla Classification Percentages 76

MOLECULAR CHARACTERIZATION OF WETLAND SOIL BACTERIAL COMMUNITIES IN CONSTRUCTED MESOCOSMS

Chapter I: Introduction

This research focused on mesocosms constructed to investigate the rhizosphere bacterial community associated with a constructed wetland at Wright-Patterson Air Force Base (WPAFB), Ohio. The wetland was built in 2000 to treat groundwater contaminated with Tetrachloroethylene (PCE) and Trichloroethylene (TCE). Twelve mesocosms were constructed to simulate the subsurface flow of the wetland, and were housed at the Wright State University (WSU) greenhouse in Dayton, OH. The mesocosm design is thoroughly explained in Chapter III of this thesis. Nine of the 12 mesocosms were planted with common wetland plants used in the constructed wetland, and three unplanted mesocosms served as controls. Three mesocosms were planted with *Eleocharis erythropoda* (Spike Rush), two were planted with *Carex comosa* (Bearded Sedge), and four were planted with *Scirpus atrovirens* (Green Bulrush) (Yan 2006). The initial intent was to evenly distribute the plant species over the nine mesocosms; however, due to a mistake identifying the plants during their collection, the distribution was not even.

The need for less expensive and more efficient remediation techniques has driven a strong interest in bioremediation. Remediation using various microbial processes has been the focal point of many research projects, but little is known about the morphology and functionality of microbial consortia that perform bioremediation. In order to completely understand and control biological remediation, engineers need to understand how organisms within the system operate.

Since the vast majority of microorganism cannot be grown under isolated conditions, and therefore cannot be studied directly, this understanding and control has not yet been achieved. An estimated 1% of microorganisms have been isolated using traditional culture laboratory methods (Pace 2008, Schloss & Handelsman 2006, Kowalchuk 2002). New molecular methodologies, such as 16S rRNA gene analysis, allow examination of the elusive 99% of the uncultured organisms by examining the organisms' DNA sequence. Numerous studies of this nature have been conducted in the field or in microcosms (Grayston 1998, Kowalchuck 2002). This is the first study of its kind to apply molecular tools to the study of microbial communities in mesocosms.

Research on wetlands constructed for the purpose of water treatment is relatively new. In 1973, the first pilot scaled constructed wetland treatment system was established combining a marsh wetland, a pond, and a meadow, in series (Kadlec & Knight 1996). However, the intricate interactions and relationships between the microbial communities and the plant life in a treatment wetland have not been thoroughly examined (Stottmeister 2003).

Microbial degradation of a contaminant, such as PCE and TCE, takes place because microorganisms use the contaminant as an electron donor (carbon source) or, as an electron acceptor (oxidant). This promotes the organism's growth and ultimately its survival (Fields 2004). However, microbes do not execute degradation without outside support. Soil is the main supporting material for plant growth, which in turn provides the structure and environment for microbial growth. These three constituents work in a delicate balance toward the ultimate outcome of bioremediation, and understanding this balance is of major interest to researchers (Stottmeister 2003).

Numerous studies have concentrated on soil properties associated with different species of plants, and plant growth and survival in different soil types (Kennedy 1995, Grayston 1998, Bardgett 1999, Meithling 2000, Yan 2006, Bezemer 2006). Those studies also looked at the composition of the microbial community. All of the studies used general methods, such as substrate utilization, to identify functional groups of bacteria, and identification based on metabolic profiles, rather than molecular technologies, to determine the composition (Kennedy 1995, Grayston 1998). Still other studies characterized the effects plants species diversity has had on a particular microbial functional group, like ammonia oxidizers (Kowalchuk 2000).

Studies have characterized microbial communities in different environments based on molecular technology; however, sample sizes are typically low compared to the large sample size presented here. Borneman *et al* (1996)., surveyed the microbial diversity of an agricultural soil in Wisconsin. They used 124 DNA sequences from 16S rRNA sequences in his research, and analyzed the sequences using the Basic Local Alignment Search Tool (BLAST), described later, for his analysis. Major Ethan Bishop used 357 sequences and analyzed them using BLAST and EstimateS (http://viceroy.eeb.uconn.edu/EstimateS). EstimateS calculates diversity parameters and allowed for complete analysis of the sample sequences; however, the sample size was extremely small (Bishop 2006). Other studies have used between 100 and 686 sequences for analysis of microbial communities and their diversity (Liu 1997, McGarvey 2004, Jannsen 2006). This study used 3,099 sequences for composition analysis, and 2820 sequences for diversity parameter analysis; it is the largest known collection of sequences, or community, to date.

The software packages used to analyze the data from the 16S rRNA gene analyses were the Ribosomal Database Project (RDP) version 9.57 Classifier and Aligner programs, *PHYL*ogeny *I*nference *P*ackage (Phylip) version 3.2, and distance based operational taxonomic unit (OTUs) and richness determination (DOTUR) version 1.53. These software packages will be described in detail in the literature review section. They allowed characterization of the entire microbial community into phyla, and produced parameters that described the diversity, richness and evenness, of each community. Therefore, we were able to compare communities, and note any effect on the diversity or composition of the microbial community. This information could be used to make inferences about the makeup of the actual wetland microbial community and its remediation potential. This research provides a baseline that will be used for comparison to subsequent contaminated mesocosm research and research specifically designed to investigate the trends identified here.

Research Objectives

The primary objectives of this research were to:

1. Determine the effects of plant presence on microbial diversity and community composition.
2. Determine the effects of plant species on microbial diversity and community composition.
3. Determine the effects of subsurface flow soil depth on microbial diversity and community composition.

The results of this research help define the relationships between microbial community diversity and plant species, microbial community diversity and depth in soil that is continuously saturated with water and experiences a subsurface flow and, most importantly, determined the impact of plant presence on the microbial community. This research provides useful information for design and construction of appropriate and efficient wetlands to biodegrade PCE and TCE.

Chapter II: Literature Review

This chapter reviews the literature that supports the major objectives of this research. First, the fundamental basis of plant and microbial interactions that take place in treatment wetlands are discussed. Then, the 16S rRNA gene analysis method and its background are discussed. Finally, the software packages used in calculating the various diversity parameters used in analysis will be introduced, and their capabilities and limitations discussed.

Treatment Wetlands and Microbial/Plant Interactions

Natural wetlands filtered groundwater long before humans began constructing artificial ones (Kadlec & Knight1996; Stottmeister 2003). Constructed wetlands have been established throughout the world to clean contamination, such as PCE and TCE, since the work of Kathe Seidel in the 1960s (Stottmeister 2003). However, the intricate interactions between the microbial communities that drive the degradation and the abiotic influences in the wetland environment are not well understood. Nevertheless, it is widely accepted that the microorganisms in a wetland transform contaminants, such as PCE and TCE, into innocuous constituents (Kadlec & Knight 1996, Stottmeister 2003).

This research was intended to identify three factors that affect microbial communities in soil. Some researchers are convinced that the soil properties are the key to understanding the degradation properties of microbial communities in treatment wetlands. They hypothesize that the soil provides the environment for certain plants to grow, and, in turn, the associated microbial community can flourish (Marrs 1991 , Marschner 2001). However, studies have also shown a direct relationship between plant species and associated microbial communities, and some researchers believe that plant species do influence the associated microbial community more so than the type of soil in a treatment wetland (Grayston 1998, Meithling 2000, Bezemer 2006).

Plants that survive in a wetland environment have adapted features. The plants are able to survive in environments that are flooded at least part of the year. All plants require water for survival, but excess water is a stressor. Therefore, wetland plants have two adaptations that allow their survival in a stressed wetland environment. The first is aerenchymous plant tissues. This tissue allows transport of gases such as oxygen from the atmosphere to the root zone, or rhizosphere. The second adaptation is the generation of adventitious roots from flooded stem tissue. This allows extraction of dissolved oxygen and other nutrients for use by the plant from the surrounding environment (Kadlec & Knight 1996, Stottmeister 2003). Oxygen not used by the plant for respiration is released into rhizosphere and other parts of the root system. This forms a protective layer around root surface, which continuously counterbalances the chemical and biological oxygen demand in the soil (Stottmeister 2003). This release rate of oxygen and other nutrients is plant species specific (Kadlec & Knight 1996).

The flow of oxygen in a plant is driven by diffusion and convective processes. The types and degree of these mechanisms are specific to each plant species. Flooded soils are oxygen deprived (Stottmeister 2003); however, plants are able to provide oxygen deep into the rhizosphere. The rhizosphere is divided into two distinct regions. The endorhizosphere is the interior root zone, and the ectorhizosphere is the root's surroundings. The area where they meet is referred to as the rhizoplane, and this area is the site of the most intensive interactions between plants, soil, and microbes (Stottmeister 2003).

Since the exudates from a plant's rhizosphere have been shown to influence microbial composition and performance, it is similarly possible that microbial communities associated with different species of plant will also be influenced (Stottmeister 2003). In a constructed wetland the main role of degradation lies with the microorganisms, not the plants. However, the plants do have an effect on the associated microbial community.

In this study, the microbial communities associated with three typical wetland plants were investigated. There are numerous studies showing the properties that various plants bring to a wetland (Grayston 1998, Stottmeister 2003, Bezemer 2006). However, there are relatively few studies that examine how plants affect the detailed microbial community composition and diversity. It is generally accepted that plants increase the diversity of a microbial community; however, no one has specifically attempted an in depth study concerning this matter.

This project used mesocosms to establish microbial communities for each of three species of plants. The plants selected were *Eleocharis erythropoda*, *Carex comosa*, and *Scirpus atrovirens*. All of these plants are in the phylum Tracheophyta (vascular plants), class Angiospermae (flowering plants) and further divided into Monocotyledonae (monocots). All of the plants chosen for this project have an emerging herb growth habit, which means that most of the above-ground part of the plant emerges above the water line in the wetland. This is an important trait because emergent plants provide surface area for microbial growth (Kadlec & Knight 1996). The studies that investigated plant species' effects on soil properties noted that plants with similar growth habits and taxonomy typically produce similar soil property effects (Kadlec & Knight 1996, Bezemer 2006). Therefore, it is reasonable to assume that the microbial community associated with these similar species of plants will only differ due to a specific property of the plant's rhizosphere, and not because of an indirect effect the plant has on soil properties.

Soil Microbial Diversity and Diversity Statistics

A soil's microbial community cannot be exhaustively sampled; therefore, samples must be used to estimate the actual diversity of organisms in that environment. Diversity consists of richness and evenness. Species richness is defined as the number of different units present in a community (Nübel 1999). The classification of a unit can be taken as a species, class, or other biological level, depending on the intent of the study. For microorganisms, it is particularly difficult to define a unit. Definite criteria have not been published. However, if the unit definition stays consistent throughout a particular study, and is adequately documented, it does not become a problem in analyzing data (Hughes 2001). Evenness is considered the relative distribution of individuals among certain predefined units, such as a species. Both of these components are investigated in this project.

Diversity can be positively linked to productivity of a community. However, microbial diversity is very hard to quantify because the tested sample will be a small subset of the site's actual population. It might not be fully representative of the population at large. Nonetheless, the estimators for comparative analysis described below have been applied to the microbial world. The estimators used for this project are described in detail later in this section. The correlation of the estimators to the new molecular techniques has not been evaluated but their use does show promise (Nübel

1999). For this project, the main goal was to document the change in microbial community diversity across depth gradients, plant species, and with and without plants. To answer these questions only relative diversities are required. Therefore, the various diversity statistics were used for analysis (Hughes 2001).

16S rRNA Gene Analysis Method

Biologically defining organisms with molecular technology uses the concept of phylogeny. A molecular basis for this concept was introduced by Olsen and Woese in 1993. This concept stated that the majority of essential genes in a genome share a common heritage or evolutionary history. A gene mutates over time. Theoretically, this change can be measured; however, the original state of an organism remains unknown. Therefore, biologists assume that two versions of a gene sequence originate from the same ancestry. Their sequence difference can be measured and compared, and ultimately the relation between two sequences can be established (Woese 1987). This is referred to as an organism's evolutionary distance.

The process of selecting a gene to be used for determining evolutionary relationships can be streamlined by focusing on genes that perform a central function and are intimately involved in the cell's activity. Several genes fit this description: rRNA, RNA polymerase, elongation factor G, proton-translocating ATPases, and others (Olsen 1993). Since several genes can be used, other criteria must be considered. A particular gene must provide enough appropriate information for analysis. In most cases, the goal of these research projects is to identify the properties and makeup of a consortium of microorganisms from a particular environmental sample, such as soil. Therefore, the gene chosen must be evolutionarily linked to its relatives and be variable enough to distinguish between unique species (Woese 1987, Clarridge 2004).

rRNA is a key element of the cell's protein synthesis process, and thus is functionally and evolutionarily homologous in all organisms. In bacteria there are 3 different rRNAs: 5S which is ~120 nucleotides, 16S which is ~1550 nucleotides, and 23S which is ~3000 nucleotides (Woese 1987; Olsen 1986; Clarridge 2004). The exact nucleotide length varies in organisms, and the aforementioned lengths are averages. The 5S and 23S rRNAs were found to be inappropriate molecular tools for the analysis of microbial communities. The 5S rRNA was not long enough to provide adequate information or detail to make an accurate comparison tool (Woese 1987). The 23S rRNA was too large a molecule, and little research has been directed into using it for genetic analysis. Therefore neither has been chosen in typical research methodologies (Olsen 1986). The most widely studied gene is the 16S rRNA gene (Schloss 2006).

The 16S rRNA gene is large enough to have conserved sequences, which are identical or nearly identical in all bacteria, and variable regions. The variable regions provide distinguishing and statistically valid measurements of evolutionary distances, and thereby of "species" or other levels of classifications of bacteria (Clarridge 2004). Regions within the 16S rRNA gene are less affected by reconfiguration that occur in the genome, and maintain a highly conserved picture of the organism's evolutionary history (Olsen 1993). This is largely due to the fact that rRNA is a critical component of the cell's function.

In cases requiring detail, such as describing a new species, it is appropriate to sequence the entire 16S rRNA gene multiple times. Also for research to distinguish between specific taxa or strains, sequencing the entire gene would be most appropriate. For descriptions of microbial communities, the 16S rRNA gene is used in two basic ways. The entire ~1550 base pair (bp) length is sequenced when relatively few microbes are analyzed, or a smaller 5', 500 bp region is used when sampling larger and more diverse communities. The first 500 bp provide sufficient information and differentiation to distinguish separate organisms, thought not always to specifically denote genus and species. Furthermore, the first 500 bp region has been shown to hold a higher percentage of diversity than any other region. Clarridge *et al*. compared 100 organisms using the 1550 bp sequence or the 500 bp sequences and found the relationships to be highly similar (Clarridge 2004). Since the goal of this thesis project was to differentiate between organisms and not to identify new species, and an extremely large sample set was generated, use of the 500 bp portion of the gene was justified.

In 1977, Woese *et al*., used the rRNA gene to completely transform the nomenclature of living organisms. Traditionally, living organisms had been classified into two distinct domains: *Prokaryotae* and *Eukaryotae*. However, as molecular genetics became a more common area of research, living organisms' genomes were investigated, and the traditional nomenclature became obsolete. Woese *et al*., used the rRNA gene to classify living organisms into three new classifications called urkingdoms. The first was the urkingdom *eubacteria*, which includes all typical bacteria. The second was *urkaryotes*, which was defined by the 18S rRNAs of the eukaryotic cytoplasm. Both of these corresponded nicely to the traditional groupings of *Prokaryote* and *Eukaryote*. However, a third classification was also introduced. The *Archaebacteria* appear to be no more related to the typical bacteria as they are to eukaryotes. Investigating the genetic makeup of organisms has unlocked an entirely new classification system (Woese 1977).

16S rRNA gene analysis was chosen as the appropriate molecular tool for the mesocosm study in this thesis. The steps in this analysis are fairly straightforward: first DNA extraction from mesocosm soils, second Polymerase Chain Reaction (PCR) to find 16S rRNA sequences within the DNA extract, third cloning of the amplified 16S rRNA products, next sequencing of the products, and finally comparative analysis of the retrieved sequences (Bishop 2006). The sampling methodology is explained in greater detail in the next chapter and by Bishop (2006). A full and detailed summary of the PCR method used is included in Appendix A. The PCR reactions generate a heterogeneous mixture of 16S rRNA sequences. It is therefore necessary to clone individual molecules in order to isolate them for sequencing. This step had the added benefit of ensuring adequate concentrations of high-quality DNA. The exact procedures for all processes are explained in the next chapter and the appendices.

The choice of appropriate primers to amplify the ~500 bp, 5' section of the 16S rRNA gene was highly dependent on the project's research goals. In this project, the goal was to identify and differentiate as many bacteria as possible from the mesocosm soil samples. Therefore, primers constructed from the conserved regions at the beginning of the gene and at the ~540 bp region were used (Clarridge 2004). These primers are often referred to as "universal" because they are built from the conserved regions that all bacteria have. However, no primer can be designed to completely anneal to all bacteria since there is variability between bacteria and other organisms (Baker 2003). The "universal" primers used in this project introduce bias into the results, because they are designed to anneal to bacteria 16S rRNA, but can anneal to genes from other organisms that are not within the domain *Bacteria*. Furthermore, they may not anneal well to the 16S rRNA genes of some bacteria. This will be discussed further in the Methodology section of this thesis.

RDP and Alignment

RDP provides ribosome related data and services to the scientific community, including online data analysis and aligned and annotated bacterial small-subunit 16S rRNA sequences. RDP had 451,545 rRNA subunit sequences as of November 8, 2007. RDP has several functions that are available to the online user. Studies have used RDP primarily to classify sequences into phyla using its Classifier function. Nercessian *et al*., and Ben-Dov *et al*., are examples of studies which applied RDP in their analyses. Nercessian identified bacterial populations active in metabolism of C_1 compounds in the sediment of a Washington state lake. RDP classifier was used to define affiliations to known phlyogenetic groups (Nercessian 2005). Eitan Ben-Dov attempted to show the advantage of using Inosine at the 3’ termini of 16S rRNA gene universal primers for the study of microbial diversity. He used RDP Classifier to assign 16S rRNA sequences to a taxonomical hierarchy (Ben-Dov 2006).

In this project, RDP was used for three important steps. RDP was used to assist in the trimming and editing process, described in detail in Chapter III. RDP was also used to assign sequences to particular phyla by the RDP Classifier program using the 80% confidence level to a sequence in the database. Finally, RDP was used to align the sequences used in the DOTUR analysis.

This project initially had 3,099 sequences for RDP analysis. The online aligners, such as ClustalW and Alignment App, were not capable of handling this number of sequences. RDP added an aligner as a part of its services, and it was able to handle this project's data set (Cole 2003). The sequence alignment was crucial to identify regions of similarity across the entire group of sequences so that homologous residues appear in the same column of alignment. It is assumed that similar residues are descended from the same common ancestral gene, and to the extent that assumption is incorrect, the alignment, and conclusions of the analysis lose justification (Olsen 1993).

In a recent study, Wong *et al*., investigated aligner limitations. They used seven prominent aligner programs: ClustalW, Muscle, T-Coffee, Dialign 2, Mafft, Dca, and ProbCons in their investigation. They found that 46.2% of the data had one or more differing tree phylogenies depending on the aligner used. They conclude that the inconsistencies were not due to the alignment procedures but rather the processes of substitutions, insertions, and deletions that make some sequences hard to align. However, many biologists do not incorporate aligner uncertainty because they accept that their alignment procedure was carefully constructed by the provider (Wong 2008). This was the position accepted in this research.

Comparative Analysis and Software

Once the alignment was completed, richness parameters and evenness were calculated, based on the evolutionary distance between the sequences. Evolutionary distances were determined using a program called Phylip, version 3.2, which was introduced in an online form in mid-1995. This package had several functions, but most importantly, it had the ability to compute evolutionary distances between nucleic acid sequences and form a distance matrix through its DNADIST function using the Jukes cantor method (Felsenstein 2005). In Chapter 12 of Bioinformatics Methods and Protocol, edited by Misener and Krawetz, Retief calls Phylip an extensive tool that covers every method of phylogenetic analysis up to 1999 (Retief 1999). A study by McGlynn *et al*., describes using Phylip to determine if distinct evolutionary pathways of tumors exist over time (McGlynn 2002). Even with the many tools Phylip has to offer, some of its components are becoming obsolete. The DNADIST tool is not obsolete, and is still in widespread use.

Calculations of richness parameters and evenness involving large sequences such as the one constructed for this project, become complicated very fast; therefore, algorithm-based software packages that perform the calculations become critical. In 2004, a program called DOTUR was introduced to overcome some of the limitations of Phylip's obsolete programs (http://www.plantpath.wisc.edu/fac/joh/dotur.html). DOTUR used an input of a distance matrix created by Phylip DNADIST program, and assigned input sequences to operational taxonomic units (OTUs) for various evolutionary distance levels using different clustering algorithms. OTUs are basic groupings determined by sequence similarity. The program calculates several known diversity indices and rarefaction data (Schloss 2005). Several studies have used DOTUR to calculate diversity parameters for data (Francis *et al.*, Sogin *et al.*). This project used DOTUR version 1.53, executed in November 2007, to calculate ACE and CHAO 1 estimators, components needed for evenness calculation, and rarefaction data.

DOTUR can use several methods to determine sequence similarities and to group sequences into OTUs according to evolutionary distances. The first method is referred to as the Nearest Neighbor method, which assumes that each sequence within an OTU is at most X% different from the most similar sequence in the group. The second method is referred to as the Furthest Neighbor method, which assumes that each sequence within an OTU is at most X% different from the any other sequence in the group. As the distance is increased the sequences added to the OTU must be within the distance from all other sequences already in the OTU. The last method that DOTUR uses is the Average Neighbor method, which is an average of the other two methods. The DOTUR manual recommends the Furthest Neighbor method for 16S rRNA gene analysis (Schloss 2005). DOTUR provides 23 output files. Each file provides information to graph rarefaction data, diversity estimators, replicate data, or other classification data useful to researchers.

As previously mentioned DOTUR groups sequences into OTUs based on their DNA sequence. There exists much controversy over the evolutionary distance levels that coincide with the species, genus, and phylum levels. No firm cutoff has been established. However, several prominent researchers have proposed: >97% similarity relates to the species level, >95% relates to the genus level, >90% relates to the family level, and >80% relates to the phylum level (Schloss 2005, Bond 1995, Everett 1999). Therefore if a sequence is >97% similar to another sequence, the organisms from which the sequences originated are then accepted to be the same species. This project uses the aforementioned cutoff values to correlate to species and phylum respectively.

DOTUR generates outputs that enable calculation of several parameters of interest. As mentioned previously, evenness is considered the relative distribution of individuals among certain predefined units, such as a species. There are numerous ways to determine evenness. This project used the popular Pielou formula for evenness calculation. The Pielou formula is the ratio of the Shannon index and the maximum value of observed OTUs when only one individual occupies each OTU (Kennedy 1995).

Good's coverage was first introduced and defined by I.J. Good in 1953 as an indication of sampling effort. Good defined coverage (C) by the following formula: C= $1-\frac{n_1}{N}$ (Good 1953). N is defined as the community size and n_1 is defined as the number of phylotypes appearing only once. Kemp and Aller described Good's coverage as a "non-parametric estimator of the proportion of phylotypes in a community of infinite size that would be represented in a smaller community" (Kemp 2004). This parameter is presented as a percentage; therefore, the higher the percentage, the higher the coverage, or sampling effort, for that particular community.

DOTUR also produces an output file entitled Rarefaction. This file has the rarefaction data for various evolutionary distances. A rarefaction curve compares observed richness, or number of OTUs, with sampling effort. The data results from averaging randomizations of the observed accumulation curve (Hughes 2001), a count of the number of OTUs at a given sampling point. Constructing rarefaction curves for the various subgroups provides a comparison of richness that was easy to interpret. DOTUR uses 10,000 randomizations in its calculations. The data can then be graphed for further analysis (Schloss 2005).

A non-parametric estimator was defined by Chao in 1984. Chao1 estimates the species total richness by the formula:

$S_{CHAO1} = S_{obs} + \frac{n_1^2}{2n_2}$, where S_{obs} is the number of observed OTUs, n_1 is the number of singletons, or OTUs occurring only once, and n_2 is the number of doubletons, or OTUs occurring twice (Hughes 2001, Schloss 2005, Chao 1984). This estimator is particularly useful when data sets are skewed toward the low-abundance classes, as they are likely to be in microbial communities (Hughes 2001). The DOTUR program uses the above formula to calculate the Chao 1 file only when $n_1=0$ and $n_2 \geq 0$. However, when $n_1>0$ and $n_2 \geq 0$ and when $n_1=0$ and $n_2=0$ DOTUR uses the formula: $S_{CHAO1} = S_{obs} + \frac{n_1(n_1-1)}{2(n_2+1)}$.

The ACE estimator incorporates data from all OTUs with fewer than 10 individuals. This includes more than just the singletons and doubletons. The ACE estimator is defined by DOTUR as the formula:

$S_{ACE} = S_{abund} + \frac{S_{rare}}{C_{ACE}} + \frac{n_1}{C_{ACE}}\gamma_{ACE}^2$, where C_{AE}=1-$\frac{n_1}{N_{rare}}$ (coverage),

$\gamma_{ACE}^2 = \max[\frac{S_{rare}\sum_{i=1}^{10} i(i-1)n_i}{C_{ACE}(N_{rare})(N_{rare}-1)} - 1, 0]$ (coefficient of variation), where

n_i is the number of OTUs with i individuals, S_{rare} is the number of OTUs with 10 or fewer individuals, S_{abund} is the number of OTUs with more than 10 individuals (Schloss 2005). Both the ACE and the Chao 1 estimators underestimate true richness at low sample sizes (Hughes 2001).

<u>Error</u>

DOTUR calculates not only the parameters but also a 95% confidence interval for some of those parameters. Typically, in statistics the confidence intervals are an equal amount both above and below the estimated mean of the parameter. DOTUR values tend to overestimate the high confidence range. The manual does not address this phenomenon. However, due to the fact that the majority of parameters estimated are proven in literature to be underestimates of richness, it is possible the DOTUR creators put more emphasis on the high confidence limit to get a more realistic range of the true estimate (Hughes 2001; Kemp & Aller 2004). Nevertheless, the error introduced by the DOTUR system, where provided, was used throughout all the subsequent calculations. Error bars often appear figures in peer reviewed articles; however, their interpretation is often incorrect. In this case, the 95% confidence intervals are used. Therefore, an overlap of more than half an error bar arm from one data set to the next indicates the data sets are not significantly different. Any overlap of less than half of an error bar arm or no overlap indicates the data sets are statistically different (Cumming 2007).

Another phenomenon typical in statistics is that confidence intervals get more refined as the sample size increases. This is due to the fact that typically confidence intervals are calculated by taking the ratio of variance to the square root of sample size as a major component of the calculation. A set of data usually has a better estimate of variance as the sample size increases so the total interval will decrease (McClave *et al.* 2008). However, in microbial analysis the variance does not follow this typical trend. For instance, in this research's data the total population was so diverse that the sample size was inadequate to estimate a variance. As more samples were taken, the variance also increased right along with sample size. This trend was seen throughout the analysis. The confidence intervals did not get smaller with increased sample size. This again was a testament of the vastness of the diversity in microbial communities.

Statistical Analysis

In order to compare microbial communities at the phylum level, as they were established by RDP, Analysis of Similarity (ANOSIM) tests were used. These tests use the ecological distances among untransformed samples from the data represented using Bray-Curtis (Clarke 1993). A random and observed test statistic, R, was generated using Primer-E v. 6.0. Data were to be statistically different if less than 5% of the generated test statistics were less than the observed test statistic. This method has recently been applied to microbiological studies (Isenhouer 2007). These tests allow some semblance of statistical integrity into studies characterizing microbial community composition.

Chapter III: Methodology

Experimental Overview

Since its construction in 2000, many research projects have focused on the groundwater treatment wetland at WPAFB, both hydraulic and remediation properties. This specific project continued the research of Major Ethan Bishop, who provided the experimental foundation summarized in the next section (AFIT/GES/ENV/06J-01).

In 2005, mesocosms were constructed at Wright State University from soil taken from both the constructed and Valle Green wetlands in Beavercreek, Ohio. The constructed wetland had already shown PCE degradation; therefore, soil from the constructed wetland was used to "inoculate" the soil from Valle Green. This ensured the soil microbial community would have a healthy consortium of PCE degraders, since, at the time, it was uncertain whether PCE degraders were part of the microbial community of Valle Green. Prior to the construction of the mesocosms, samples from the inoculated soil were taken to establish baseline data for the microbial community prior to planting of the columns or PCE exposure.

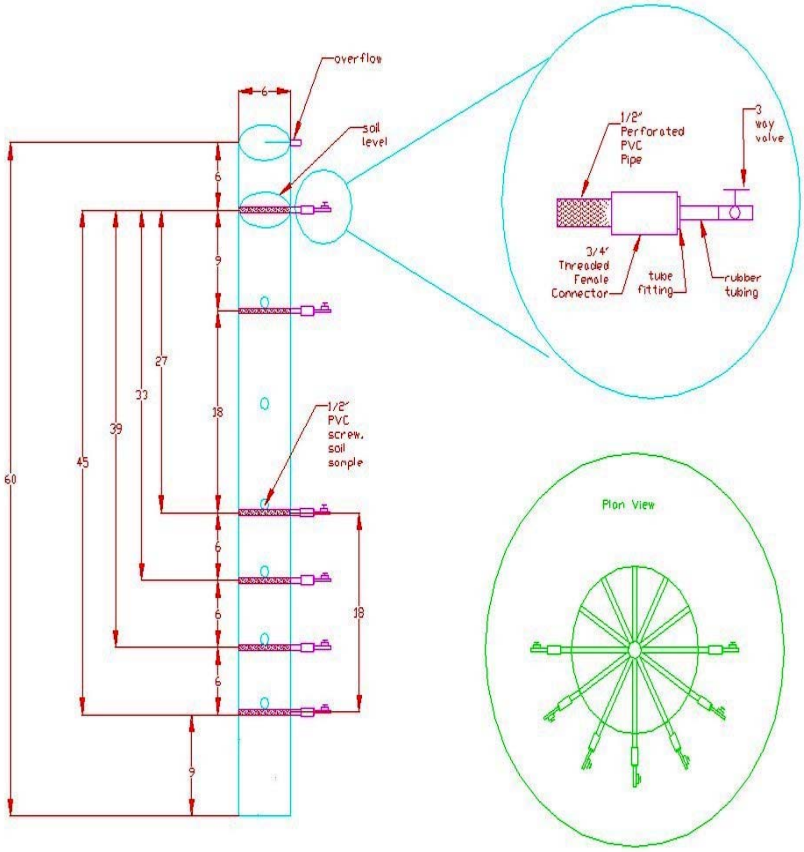


Figure 1: Mesocosm Design
All measurements in inches

Figure 1 illustrates the column design and dimensions for the mesocosms (Bishop, 2006). Each mesocosm was constructed from 6-in diameter PVC pipe with a depth representative of the actual WPAFB constructed wetland. Three wetland plants, *Eleocharis erythropoda* (Spike Rush), *Carex comosa* (Bearded Sedge), and *Scirpus atrovirens* (Green Bulrush), were used in this experiment. A single species was planted in each mesocosm in an effort to characterize its effects on its associated microbial community. Three control mesocosms were also established for comparison of microbial communities that developed without higher plant association.

Table 1: Mesocosm Plantings (Bishop 2006)

Mesocosm	Species
1	*Carex comosa*
2	*Carex comosa*
3	Control
4	*Eleocharis erythropoda*
5	*Scirpus atrovirens*
6	*Scirpus atrovirens*
7	*Eleocharis erythropoda*
8	Control
9	*Scirpus atrovirens*
10	*Eleocharis erythropoda*
11	Control
12	*Scirpus atrovirens*

After the plants grew for 2 months, 5 gram soil samples were taken from each mesocosms at each of three separate depths: depth 1, 49 inches (bottom sample), depth 2, 31 inches (middle sample), and depth 3, 13 inches (top sample). Root mass was observed in all samples demonstrating that the plant roots had extended the entire length of the mesocosms (Bishop 2006).

DNA was extracted from the 36 soil samples using the Mo Bio PowerSoil^{TM} DNA Isolation Kit with the standard protocol (Appendix C). PCR was performed with these DNA extracts as the templates to amplify the 16S rRNA genes. Universal primers, E8F and E533R, were used for PCR because they are both very sensitive to detection of bacteria. While primer E8F has a slight affinity for Archaea and primer E533R has an affinity for both Archaea and Eukarya, these two universal primers are specific enough to bacteria to meet the goals of this project (Baker 2003). The PCR protocol and conditions used for this experiment are summarized in Appendix A. Of the PCR products generated, 357 were cloned and sequenced during the course of the Bishop project. The original PCR reactions were frozen at -20ºC for future research (Bishop 2006).

Nomenclature

This project combined data from Bishop's research with new sequence data taken from Bishop's original PCR reactions that had been stored as described above. Therefore, a unique nomenclature was required. Bishop labeled all his soil samples with an "A" and two subsequent numbers. The "A" represented August, the month of soil extraction; 1st number depicted the column number; and the 2nd number represented the depth of the sample. During the course of generating the sequenced data, additional numbers were added to the sample name. The subsequent numbering represented the cloning reaction, plate number and colony number respectively.

As new cloning reactions were performed for this project, the labeling system was adjusted to differentiate the Bishop data from the new data. The first letter represented the month of cloning (Appendix B). The next letter was always "L", illustrating that the cloning reaction was performed during the Leon project. The number after the "L" was the cloning reaction. This project performed only one cloning reaction for each PCR tube, therefore, the number after the letter "L" was always 1 for all the new data. The subsequent numbers represented the plate number and colony number respectively. On average, five plates were used for each cloning reaction. For instance, the sample identified as Ju53.L1.1.1 is a sample that was cloned in the month of June, from column 5, depth 3, it is a Leon first cloning, and it was the first colony picked from plate 1. The detailed nomenclature was crucial to this project. The column and depth a particular sample originated from was used throughout the analysis of all the data. During the sequencing several sample names had to be adjusted due to space limitations and procedural criteria. Therefore the original nomenclature was not entirely preserved. However each sample is uniquely identifiable, and the column number and depth were always evident.

Laboratory procedures

PCR amplifications from the Bishop project were frozen and stored at -20ºC. In January of 2007, Bishop's stored PCR products were used for additional cloning and DNA sequencing. The cloning was executed using the *StrataClone^{TM} PCR Cloning Kit (*Stratagene*,* La Jolla, CA; Appendix D*)*.

Four to five plates of Luria-Bertani (LB) media, supplemented with ampicillin (AMP), were used for each cloning. Each plate received on average 50 µl of the transformation mixture. LB media is a rich medium commonly used to grow *E. coli*, and 1L is prepared using the following recipe (Difco Manual 1998):

- 10.0 g Tryptone
- 5.0 g Yeast Extraction
- 10.0 g NaCl
- Distilled or deonized water, used to fill to 1 Liter
- Adjust the pH to 7.5
- 15.0 g of agar

After LB media was thoroughly mixed, it was autoclaved on liquid cycle for 20 minutes at 15 psi and 121ºC. Next, the mixture was placed in a 55 ºC water bath to cool. AMP was added to a final concentration of 50 µg/ml. The addition of AMP to the media was a crucial step to activate the selectable marker built into the standard cloning kit. Also the substrate 5-bromo-4-chloro-3-indolyl-β-D-galactopyranoside (X-gal), from a stock concentration of 20 mg/ml was diluted to a final concentration of 40 µg/ml in the medium for blue-white screening (Chaffin 1998). The purpose of AMP and X-gal addition is explained in a later section.

The plates onto which transformations from the Strataclone kit had been spread were incubated overnight at 37ºC. 100 white colonies from each transformation were chosen from the plates and aseptically transferred with sterile toothpicks to a Falcon® tube with 5 mL of LB broth with AMP (final concentration of 50 ug/ml). AMP in this media helped maintain selection for cells that received a plasmid. Following ~16 hour incubation at 37ºC with shaking at 150-175 rpm, the Falcon® tubes were centrifuged at 6,800 x *g* with an Avanti® J-26 XPI centrifuge for 15 minutes at 20ºC. Media was poured off, the tubes were blotted on paper towels, and cell pellets were used for plasmid isolation. QIAgen's QIAprep Spin Miniprep Kit (QIAgen Inc., Valencia, CA) was used to purify and isolate plasmid DNA. The QIAprep Spin Miniprep Kit Using a Micro centrifuge protocol was used for this procedure (Appendix E). Throughout the process the samples were labeled uniquely.

Quality Check for Laboratory Procedures

During the laboratory procedures numerous quality checks were in place. The plasmids and competent cells used in the Strataclone kit were engineered with several verification vehicles. PCR products were cloned into a plasmid which would replicate within a host *E.coli* cell. The intention was that only plasmids within a cell that had the PCR product inserted into them would be able to replicate. It was possible that the cloning procedures produced plasmids, and ultimately cells, that were replicating without the PCR product insert. Therefore blue/white screening and a selectable marker were used. These procedures are explained below.

AMP is an antibiotic used to prevent contamination; however, that was not its primary purpose in this procedure. Cells that received a cloning plasmid were resistant to AMP and could grow uninhibited on the LB+AMP media. Another goal was to only proceed with cells that received a plasmid with a PCR product insert. Blue-white screening is a useful tool to make this determination. A successful cloning disrupts an enzyme reaction within the cell. X-gal is colorless modified galactose sugar, and is the substrate for this reaction (Chaffin 1998, Stratagene® 2007). If a PCR product has been inserted into the functional gene encoding the enzyme, the XGAL will not be used by the cells, and the resultant colony will be white on the plate (Messing 1977, Stratagene® 2007). The cells that do use the XGAL, indicating that they carry a plasmid with no PCR insert, will turn blue. The white colonies were removed from the plate and placed in 5 ml of LB broth with AMP. The AMP here maintains the selection of cells that have the plasmid because it is possible for the cells to lose the plasmid during growth.

***Eco*R1 Restriction Enzyme Digestion and Gel Electrophoresis**

Once the plasmids were isolated, quality checks were run on selected samples to ensure that the correct plasmids had been isolated and that they had the inserted PCR products prior to sequencing. After four cloning reactions in which the insertion was 100% efficient, this particular step was no longer performed, to expedite the sequencing process. Isolated plasmids were digested with the restriction enzyme *Eco*R1, which cuts the plasmid at sites that flank the PCR insert. Figure 2 below illustrates a gel demonstrating the successful separation of the target DNA. The PCR insert bands migrate to approximately the 500 bp band, while the plasmid band is approximately 3.5 kb. The variability in the migration of the PCR band in the different lanes was expected since the organisms may have a range of ~450 bp to ~600 bp inserts (Woese 1987). The protocol used for the restriction digest is summarized in Appendix F, and all gels are shown in Appendix G.

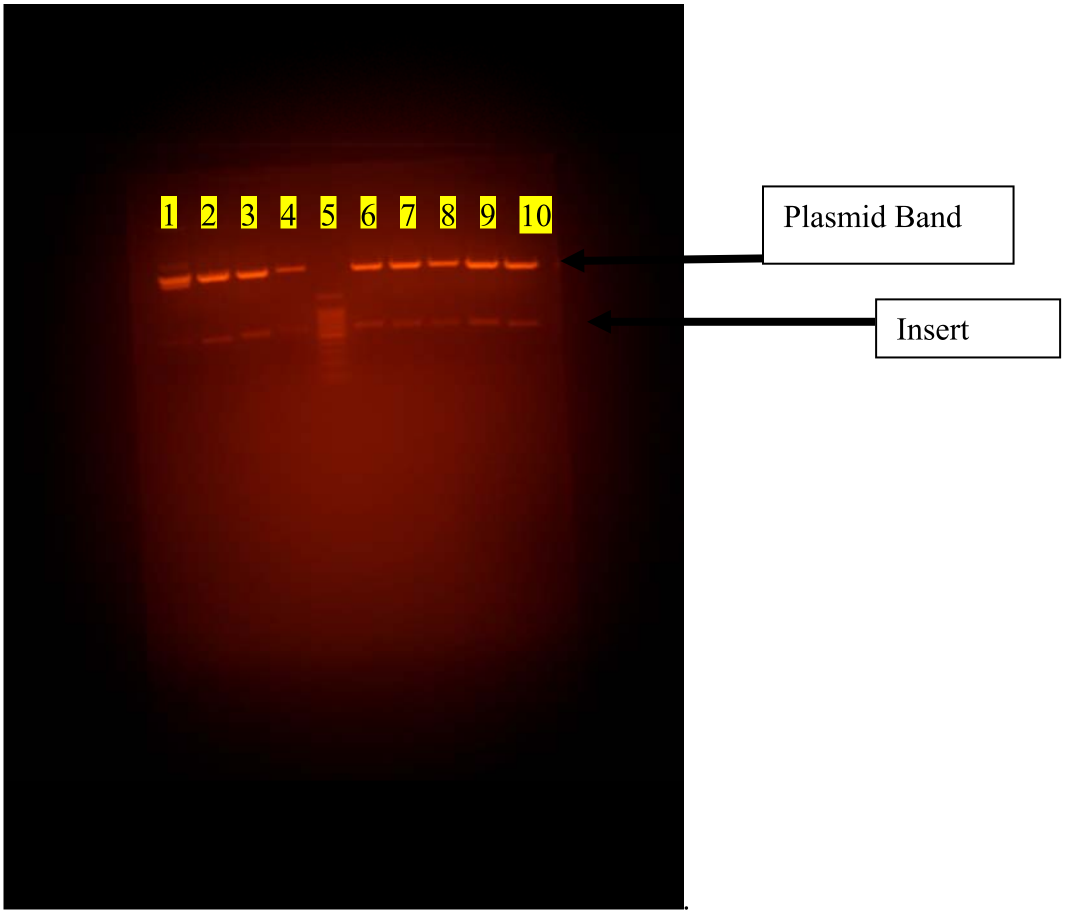


Figure 2: Gel from Ap53.L1
Lane 1-Ap53.L1.5.6; Lane 2-Ap53.L1.3.18; Lane 3-Ap53.L1.3.14; Lane 4-Ap53.L1.3.10; Lane 5: 100bp ladder; Lane 6-Ap53.L1.2.5; Lane 7-Ap53.L1.3.2; Lane 8-Ap53.L1.5.18; Lane 9-Ap53.L1.5.14; Lane 10-Ap53.L1.5.13

Sequencing and Trimming

Through the quality control procedures described above, it was evident that the cloning and plasmid purification protocol worked and PCR inserts could be sequenced. Prior to sequencing, DNA concentrations were determined because both facilities required a concentration of 50 ng/µl or above for sequencing. The sample DNA concentrations were determined after the plasmid purification and isolation by a nanodrop system. This system is a spectrometer that evaluates samples as small as 1µl. The DNA samples were loaded onto the nanodrop machine and DNA concentrations were recorded by hand for the sequencing facilities. Only samples that fell within the desired range were submitted for sequencing.

Due to the large number of isolated plasmids, sequencing was handled both at the WSU Genomics Laboratory (EEEGL) and through the Ohio State University's (OSU) Plant-Microbe Genomics Facility (PMGF). The EEEGL used a Beckman-Coulter CEQ8000 Genetic Analysis System, while the PMGF utilized an Applied Biosystems platform. Both facilities used the M13F primer to recognize the Strataclone plasmid in sequencing reactions, and provided output data in FASTA format. Chromatograms were also included for the data. On a few occasions, samples that failed to sequence at the EEEGL were submitted to the PMGF, which returned positive results for those samples. This prompted a closer look at the sequences from the two laboratories. Although both laboratories produced useable sequences for analysis, the PMGF yielded readable sequence output for 99% of plasmids submitted, whereas EEEGL produced usable sequences an average of 90% of the submissions. Sequences from the PMGF were typically longer (over 600bp), also.

A thorough quality check procedure ensured only good quality sequence data were further analyzed. As a first step, all sequences less than 300 base pairs (bp) were automatically omitted, because they did not provide a large enough region of the 16S rRNA gene to provide valid contribution to the project. During identification and deletion of sequences with less than 300 bp, sequences with numerous N's or repeated letters were identified and highlighted.

Repeated letters in sequences indicated possible contamination of the sample. N's appear in place of nucleotides when insufficient evidence was picked up with the sequences. The N's indicate a point where any nucleotide could have matched the sequence analysis. Numerous N's indicates that the sample was not concentrated enough to produce a valid sequence (Isenhouer 2008, Servaites 2007). A qualitative assessment of these sequence's chromatograms was performed based on background noise and peak height and spread. This step helped to identify samples that were contaminated or sequenced at low concentrations and those sequences were omitted. An example of this step of editing is summarized in Figure 3.

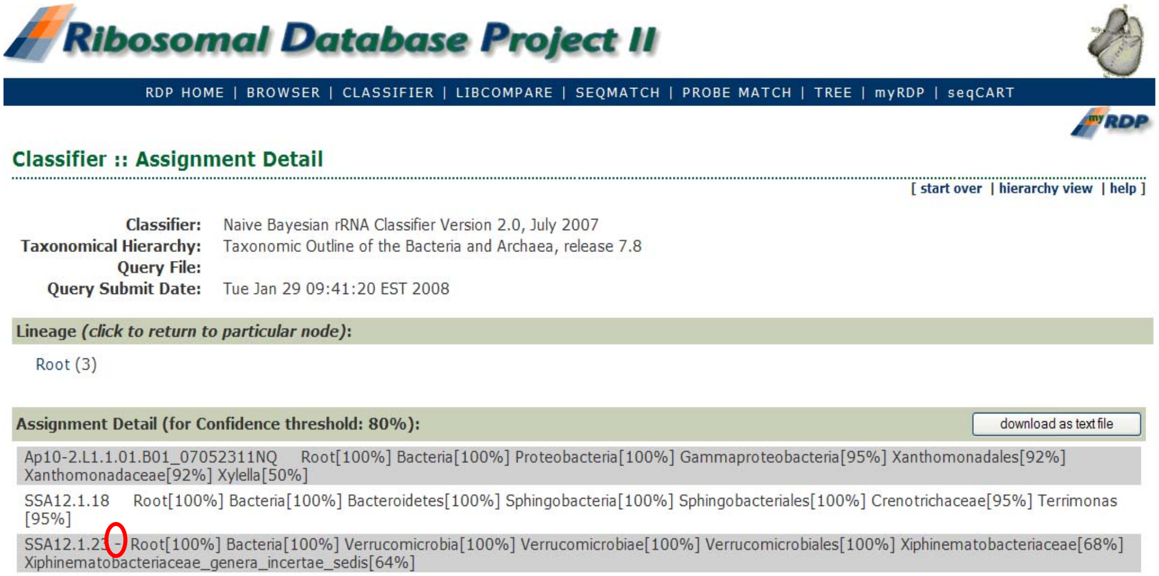
```
>Ap10-2.L1.1.01.B01_07052311NQ    768     14    768   CEQ
GGAGTTGTTCACACGGGCCAGTGAGCGCGCTAATACGATCTCACTATAGGGCGAATTGGAGCTCCCGCGTTGCCACGCT
ACTAGAACTAGTGGATCCCCCGGGTCTTGCAGCACATTGTTGGAATTCGCCCTTAGAGTTTGATCCTGGCTCAGAGTGA
ACGCTGGCGGCAGGCTAACACATGCAAGTCGAACGGCAGCACAGGGGAGCTTGCCTNGGGTGGCGAGTGGCGGACGG
GTGAGGAATACATGGGAATCTACCCTGTCGTGGGGGATAACGTAGGGAAACTTACGCTAATACCGCATACGACCGAGA
GTTGAAAGCGGCGGACCGAAGGCGTCACGCGACTGGATGAGCCCATGTCGGATTAGCTAGTTGGCGGGGTAAAGGCCC
ACCAAGGCGACGATCCGTAGCTGGTCTGAGAGGATGATCAGCCACACTTGGAACTGAGACACGGTCCAAGACTCCTAC
GGGAGGCAGCAGTGGGGAATATTGGACAATGGGCGCAAGGTATCCCAGCCATGCCGCGTGGGTGAAAGAAGGCCTT
CGGGTTGTAAAGCCTTTTGTCCCGGAAAGAAAAGCACGGGATTAAATACCCTCGTGTGATGACGGTACCCGGAAGAA
ATACGCAACCGGCTACCTTTCGTGTCAAGCAGCCCCCGGTTCAAAAGGGCGAAAATCCCACAAGTTGGAATATTCAAG
GCCTAATCGGATAACCGTCGACCCTCGAGCGCGCGGGCCCGGTTACCAAGCCTTTTTGTTTCCCTT
>SSA12.1.18
GGCCAGTGAATTGTAATACGACTCTTCTTATAGGGCGAATGGGGCCCTCTAGATGCTGCTCGAGCGGCCGCCAGTGTGA
TGGATATCTGCAGAATTCGCCCTTAGAGTTTGATCCTGGCTCAGGGGATGAACGCTAGCGGCAGGCTTAATACATGCAA
GTCGTGGGGCAGCATGTCCCGCAGCAATGCGGGATGATGGCGACCGGCAAACGGGTGCGGAACACGTACACAACCTTC
CTTTTAGTGGAGAATAGCCCAGGGAAACTTGGATTAATACTCCGTAACATATAAGAAGTGGCATCACTTTTATATTAAA
GCAGCAATGCGCTGGAAGATGGGTGTGCGGCTGATTAGATAGTTGGCGGGGTAACGGCCCACCAAGTCGACGATCAGT
AACTGGTGTGAGAGCACGACCAGTCACACGGGCACTGAGACACGGGCCCGACTCCTACGGGAGGCAGCAGTAAGGAA
TATTGGTCAATGGACGCAAGTCTGAACCAGCCATGCCGCGTGGAGGATGAAGGTCCTCTGGATTGTAAACTTCTTTTAT
TTGGGAGGAAATCCATTTTTTCTAAAATGGTTGACGGTACCAGATGAATAAGCACCGGCTAACTCTGTGCCAGCAGCCC
CGGTCAAAGGGCGAATCCAGCACACTGGCGGCCGTTACTAGTGGATCGAGCTGGTACAAGCTGGCGTAATATGGCATG
CTGTTTCGGTGTAATTGTATCGCTCCANTCCCACAACAACAGCCGAGCATAGGGTAAGCTGTGGT
>SSA12.3.17
TAAGCAAGCGCGGAGTGAAATTAGTAAATACGACTCACTATAGGGCGAATTGGGCCCTTCTAGATGCATGCTCGAGCG
GCCCGCAGTGTGATGGATATCTGCAGAATTCGCCCTTAGAGTTTGATCCTGGCTCAGAACGAACGCTGGCGGCGTACTT
AACACATGCAAGTCGAACGAGAAAAGAGACTTCGGTCTCCGAAGTAAAAGTG
>SSA12.1.23
GCCAGTGAATTGTAATACGACTCACTATAGGGCGAATTGGGCCCCTCTAGATGCATGCTCGAGCGGCCGCCAGTGTGAT
GGATATCTGCAGAATTCGCCCTTTGACCGGGGCTGCTGGCACAGAGTTAGCCGTCTCTTCCTCTTGCGGTACTATCACTT
GCTTGTTCCCCGCATGACAGGAGTTTACAACCCGAAGGCCTTCATCCTCCACGCGGCGTCGCTCCATCAGGGTTTCCCC
CATTGTGAAAAATTCTCGACTGCTGCCACCCGTAGGTGTCTGGACCGTATCTCAGTTCCAGTGTGGCTGGTCGTCCTCTC
AGACCAGCTACCCGTCATCGCCATGGTGGGCCGTTACCCCGCCATCTAGCTGATAGGCCGCGAGCTCATCAGGAAGCG
CATTGCTGCTTTGGCTTTTCCTCCAATCGAAGGATGGCCATATGCGGTATTAATTCGCCTTTCGGCGAGCTATCCCCCAC
TTCCCGGCAGATTGCTCACGTGTTACGCACCCGTGCGCCACTGAACCAAGCCTGTATTGCTACAAACCTAGTCCGTTCG
ACTTGCATGTCTTATCCACGCCGCCAGCGTTCGTTCTGAGCCAGGATCAAACTCTAAGGGCGAATCCAGCACACTGCGG
GCGTACTAGTGGATCGAGCTCGGTACAGCTGCGTATCA
>F11.L1.3.33.F07_070412218E    680     0    680   CEQ
CGGCCAGTGAGCGCGCTAATACGACTCACTATAGGGCGAATTGGAGCTCCCGCGTGCCCGCTACTAGAACTAGTGGAT
CCCCCGGGACTGCAGCAATGGTGGAATTCGCCCTTAGAGTTTGATCCTGGCTCAGATTGAACGCTGGCGGCAAATGGCA
TAATAAAAACAAACAAATGGACAAAAAAGNTACAGAAAAAACGGCNGAAAAAAAAGGGGGGGGGGGGCAAAAAAC
CACAAAAAAAAAGGGTAAAAGGAAGGGTTGGGGCCGGAAAAAACGGGGGNGGGGTGGAAAGGTTAAAAAAAATTAAA
ACAAAATTTTCCCCGGGGGGGAAAAAAAAAAACCGGGGTTTTTTTGGGCCACACAACACCCCCACCCACAAAAAAAAAT
TTTTTTTTTTTTTTTAAAAAAAAAAAGGGGGGGGTGTTTTTTTTTGGGGGGGGGGGGGGGGGGGGGGGGGGGAAGGG
GGGGGGGGGGGGGGGGGGGGAAAAAAAAAAAACACACCACAAAAAAAAAAAAAAAAAAAAAACACCCCCCCCCCC
CCCCCCCCCCACCCCTCACTTTTTTTTTTTTCTCCCCCCCCCCCCGCCGCGNGGGGGGGGGGGGGGAAAAAAAAAAAAA
AAAGAAGGGGGGGGGGGGGTAAAAAAAAAAAAAAAAAAAAAAAAAATTTTTTTATATATATT
```

Figure 3: Editing Step 1
Example of short sequences and sequences with repeated letters and N's

In the next step, sequences were analyzed by the Ribosomal Database Project II release 9.57 (RDP) Classifier system to determine the closest match to known 16S rRNA sequences within the RDP database. Each rRNA query sequence was assigned to a phylum at an 80% confidence match to a sequence within the database. An average of 0.5% of the sequences fell into an Unclassified Root category (Cole *et al*. 2007).

Unclassified Root refers to sequences for which the Classifier cannot identify as bacterial 16S genes. They could have been non 16S genes, or 16S genes from non bacteria, or sequences of low quality (RDP Staff 2007). The Unclassified Bacteria category referred to any sequence that was identified as Bacteria but did match particular phyla with a confidence level of 80% or better. Pie graphs were constructed for each community based on the RDP Classifier program results.

The symbol "-"after a sequence in the assignment detail view of the RDP Classifier program indicated that the match occurred using the reverse complement of that particular sequence (Cole 2007; Wang 2007). The sequences were identified and reverse complemented (RC) using the Reverse Complement Program (http://www.bioinformatics.org/sms/rev_comp.html). An example of this step of editing is summarized in Figure 4. This was done so that the sequences would be in the proper orientation (reading 5' to 3') prior to the steps described below, which were a continuation of the editing and trimming quality control process.


Ribosomal Database Project II
RDP HOME | BROWSER | CLASSIFIER | LIBCOMPARE | SEQMATCH | PROBE MATCH | TREE | myRDP | seqCART
myRDP
[start over | hierarchy view | help]
download as text file

Classifier :: Assignment Detail

Classifier: Naive Bayesian rRNA Classifier Version 2.0, July 2007
Taxonomical Hierarchy: Taxonomic Outline of the Bacteria and Archaea, release 7.8
Query File:
Query Submit Date: Tue Jan 29 09:41:20 EST 2008

Lineage ***(click to return to particular node)*****:**

Root (3)

Assignment Detail (for Confidence threshold: 80%):

Ap10-2.L1.1.01.B01_07052311NQ Root[100%] Bacteria[100%] Proteobacteria[100%] Gammaproteobacteria[95%] Xanthomonadales[92%] Xanthomonadaceae[92%] Xylella[50%]

SSA12.1.18 Root[100%] Bacteria[100%] Bacteroidetes[100%] Sphingobacteria[100%] Sphingobacteriales[100%] Crenotrichaceae[95%] Terrimonas [95%]

SSA12.1.23 - Root[100%] Bacteria[100%] Verrucomicrobia[100%] Verrucomicrobiae[100%] Verrucomicrobiales[100%] Xiphinematobacteriaceae[68%] Xiphinematobacteriaceae_genera_incertae_sedis[64%]

```
>Ap10-2.L1.1.01.B01_07052311NQ    768     14    768   CEQ
GGAGTTGTTCACACGGGCCAGTGAGCGCGCTAATACGATCTCACTATAGGGCGAATTGGAGCTCCCGCGTTGCCACGCT
ACTAGAACTAGTGGATCCCCCGGGTCTTGCAGCACATTGTTGGAATTCGCCCTTAGAGTTTGATCCTGGCTCAGAGTGA
ACGCTGGCGGCAGGCTAACACATGCAAGTCGAACGGCAGCACAGGGGAGCTTGCCTNGGGTGGCGAGTGGCGGACGG
GTGAGGAATACATGGGAATCTACCCTGTCGTGGGGGATAACGTAGGGAAACTTACGCTAATACCGCATACGACCGAGA
GTTGAAAGCGGCGGACCGAAGGCGTCACGCGACTGGATGAGCCCATGTCGGATTAGCTAGTTGGCGGGGTAAAGGCCC
ACCAAGGCGACGATCCGTAGCTGGTCTGAGAGGATGATCAGCCACACTTGGAACTGAGACACGGTCCAAGACTCCTAC
GGGAGGCAGCAGTGGGGAATATTGGACAATGGGCGCAAGGTATCCCAGCCATGCCGCGTGGGTGAAAGAAGGCCTT
CGGGTTGTAAAGCCTTTTGTCCCGGAAAGAAAAGCACGGGATTAAATACCCTCGTGTGATGACGGTACCCGGAAGAA
ATACGCAACCGGCTACCTTTCGTGTCAAGCAGCCCCCGGTTCAAAAGGGCGAAAATCCCACAAGTTGGAATATTCAAG
GCCTAATCGGATAACCGTCGACCCTCGAGCGCGCGGGCCCGGTTACCAAGCCTTTTTGTTTCCCTT
>SSA12.1.18
GGCCAGTGAATTGTAATACGACTCTTCTTATAGGGCGAATGGGGCCCTCTAGATGCTGCTCGAGCGGCCGCCAGTGTGA
TGGATATCTGCAGAATTCGCCCTTAGAGTTTGATCCTGGCTCAGGGGATGAACGCTAGCGGCAGGCTTAATACATGCAA
GTCGTGGGGCAGCATGTCCCGCAGCAATGCGGGATGATGGCGACCGGCAAACGGGTGCGGAACACGTACACAACCTTC
CTTTTAGTGGAGAATAGCCCAGGGAAACTTGGATTAATACTCCGTAACATATAAGAAGTGGCATCACTTTTATATTAAA
GCAGCAATGCGCTGGAAGATGGGTGTGCGGCTGATTAGATAGTTGGCGGGGTAACGGCCCACCAAGTCGACGATCAGT
AACTGGTGTGAGAGCACGACCAGTCACACGGGCACTGAGACACGGGCCCGACTCCTACGGGAGGCAGCAGTAAGGAA
TATTGGTCAATGGACGCAAGTCTGAACCAGCCATGCCGCGTGGAGGATGAAGGTCCTCTGGATTGTAAACTTCTTTTAT
TTGGGAGGAAATCCATTTTTTCTAAAATGGTTGACGGTACCAGATGAATAAGCACCGGCTAACTCTGTGCCAGCAGCCC
CGGTCAAAGGGCGAATCCAGCACACTGGCGGCCGTTACTAGTGGATCGAGCTGGTACAAGCTGGCGTAATATGGCATG
CTGTTTCGGTGTAATTGTATCGCTCCANTCCCACAACAACAGCCGAGCATAGGGTAAGCTGTGGT
>SSA12.1.23(RC)
TGATACGCAGCTGTACCGAGCTCGATCCACTAGTACGCCCGCAGTGTGCTGGATTCGCCCTTAGAGTTTGATCCTGGCT
CAGAACGAACGCTGGCGGCGTGGATAAGACATGCAAGTCGAACGGACTAGGTTTGTAGCAATACAGGCTTGGTTCAGT
GGCGCACGGGTGCGTAACACGTGAGCAATCTGCCGGGAAGTGGGGGATAGCTCGCCGAAAGGCGAATTAATACCGCAT
ATGGCCATCCTTCGATTGGAGGAAAAGCCAAAGCAGCAATGCGCTTCCTGATGAGCTCGCGGCCTATCAGCTAGATGG
CGGGGTAACGGCCCACCATGGCGATGACGGGTAGCTGGTCTGAGAGGACGACCAGCCACACTGGAACTGAGATACGG
TCCAGACACCTACGGGTGGCAGCAGTCGAGAATTTTTCACAATGGGGGAAACCCTGATGGAGCGACGCCGCGTGGAGG
ATGAAGGCCTTCGGGTTGTAAACTCCTGTCATGCGGGGAACAAGCAAGTGATAGTACCGCAAGAGGAAGAGACGGCTA
ACTCTGTGCCAGCAGCCCCGGTCAAAGGGCGAATTCTGCAGATATCCATCACACTGGCGGCCGCTCGAGCATGCATCTA
GAGGGGCCCAATTCGCCCTATAGTGAGTCGTATTACAATTCACTGGC
```

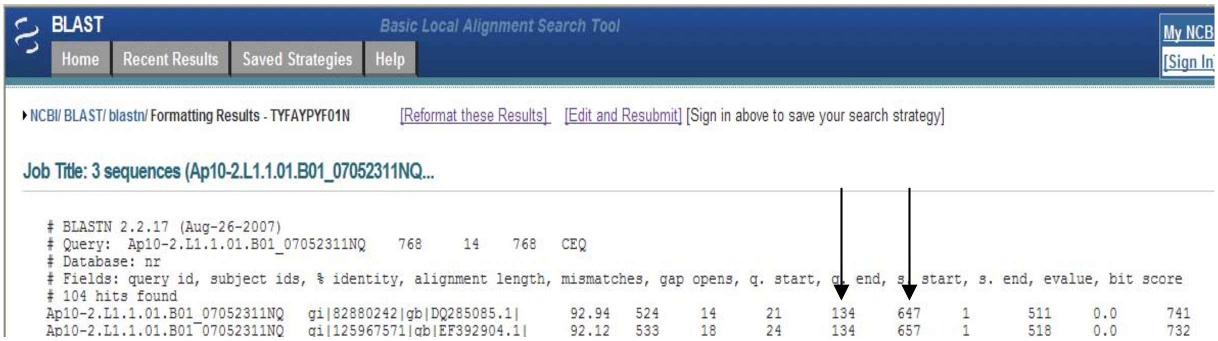
Figure 4: Editing Step 2
RDP Classifier program assignment detail view to identify RC sequences

At this stage sequences could still have plasmid, primers, and *Eco*R1 restriction sites sequences still embedded in them. The next step was to trim the sequences to remove these irrelevant pieces. This is a consequence of the sequencing reaction, whereby the DNA extension from the sequence primer could proceed past the PCR insert of interest, and into the flanking *Eco*RI restriction sequences and further plasmid sequences. The *Eco*RI restriction sites provided a convenient means for locating these flanking sequences, as were the sequences of the original primers used to amplify the 16S rRNA gene. Since these sequences represented something other than the actual 16S rRNA sequences that were needed for analyses, it was important they were trimmed away. The primers and restriction sites were identified by the Microsoft Word 2003 Word Find function and highlighted. An example of this step of editing is summarized in Figure 5.

```
>Ap10-2.L1.1.01.B01_07052311NQ    768     14    768   CEQ
GGAGTTGTTCACACGGGCCAGTGAGCGCGCTAATACGATCTCACTATAGGGCGAATTGGAGCTCCCGCGTTGCCACGCT
ACTAGAACTAGTGGATCCCCCGGGTCTTGCAGCACATTGTTGGAATTCGCCCTTAGAGTTTGATCCTGGCTCAGAGTGA
ACGCTGGCGGCAGGCTAACACATGCAAGTCGAACGGCAGCACAGGGGAGCTTGCCTNGGGTGGCGAGTGGCGGACGG
GTGAGGAATACATGGGAATCTACCCTGTCGTGGGGGATAACGTAGGGAAACTTACGCTAATACCGCATACGACCGAGA
GTTGAAAGCGGCGGACCGAAGGCGTCACGCGACTGGATGAGCCCATGTCGGATTAGCTAGTTGGCGGGGTAAAGGCCC
ACCAAGGCGACGATCCGTAGCTGGTCTGAGAGGATGATCAGCCACACTTGGAACTGAGACACGGTCCAAGACTCCTAC
GGGAGGCAGCAGTGGGGGAATATTGGACAATGGGCGCAAGGTATCCCAGCCATGCCGCGTGGGTGAAAGAAGGCCTT
CGGGTTGTAAAGCCTTTTTGTCCCGGAAAGAAAAGCACGGGATTAAATACCCTCGTGTGATGACGGTACCCGGAAGAA
ATACGCAACCGGCTACCTTTCGTGTCAAGCAGCCCCCGGTTCAAAAGGGCGAAAATCCCACAAGTTGGAATATTCAAG
GCCTAATCGGATAACCGTCGACCCTCGAGCGCGCGGGCCCGGTTACCAAGCCTTTTTGTTTCCCTT
>SSA12.1.18
GGCCAGTGAATTGTAATACGACTCTTCTTATAGGGCGAATGGGGCCCTCTAGATGCTGCTCGAGCGGCCGCCAGTGTGA
TGGATATCTGCAGAATTCGCCCTTAGAGTTTGATCCTGGCTCAGGGGATGAACGCTAGCGGCAGGCTTAATACATGCAA
GTCGTGGGGCAGCATGTCCCGCAGCAATGCGGGATGATGGCGACCGGCAAACGGGTGCGGAACACGTACACAACCTTC
CTTTTAGTGGAGAATAGCCCAGGGAAACTTGGATTAATACTCCGTAACATATAAGAAGTGGCATCACTTTTATATTAAA
GCAGCAATGCGCTGGAAGATGGGTGTGCGGCTGATTAGATAGTTGGCGGGGTAACGGCCCACCAAGTCGACGATCAGT
AACTGGTGTGAGAGCACGACCAGTCACACGGGCACTGAGACACGGGCCCGACTCCTACGGGAGGCAGCAGTAAGGAA
TATTGGTCAATGGACGCAAGTCTGAACCAGCCATGCCGCGTGGAGGATGAAGGTCCTCTGGATTGTAAACTTCTTTTAT
TTGGGAGGAAATCCATTTTTTCTAAAATGGTTGACGGTACCAGATGAATAAGCACCGGCTAACTCTGTGCCAGCAGCCC
CGGTCAAAGGGCGAATCCAGCACACTGGCGGCCGTTACTAGTGGATCGAGCTGGTACAAGCTGGCGTAATATGGCATG
CTGTTTCGGTGTAATTGTATCGCTCCANTCCCACAACAACAGCCGAGCATAGGGTAAGCTGTGGT
>SSA12.1.23(RC)
TGATACGCAGCTGTACCGAGCTCGATCCACTAGTACGCCCGCAGTGTGCTGGATTCGCCCTTAGAGTTTGATCCTGGCT
CAGAACGAACGCTGGCGGCGTGGATAAGACATGCAAGTCGAACGGACTAGGTTTGTAGCAATACAGGCTTGGTTCAGT
GGCGCACGGGTGCGTAACACGTGAGCAATCTGCCGGGAAGTGGGGGATAGCTCGCCGAAAGGCGAATTAATACCGCAT
ATGGCCATCCTTCGATTGGAGGAAAAGCCAAAGCAGCAATGCGCTTCCTGATGAGCTCGCGGCCTATCAGCTAGATGG
CGGGGTAACGGCCCACCATGGCGATGACGGGTAGCTGGTCTGAGAGGACGACCAGCCACACTGGAACTGAGATACGG
TCCAGACACCTACGGGTGGCAGCAGTCGAGAATTTTTCACAATGGGGGAAACCCTGATGGAGCGACGCCGCGTGGAGG
ATGAAGGCCTTCGGGTTGTAAACTCCTGTCATGCGGGGAACAAGCAAGTGATAGTACCGCAAGAGGAAGAGACGGCTA
ACTCTGTGCCAGCAGCCCCGGTCAAAGGGCGAATTCTGCAGATATCCATCACACTGGCGGCCGCTCGAGCATGCATCTA
GAGGGGCCCAATTCGCCCTATAGTGAGTCGTATTACAATTCACTGGC
```

Figure 5: Editing Step 3
Identifying primers (yellow) and restriction sites (pink).

The sequences are then uploaded into the mega Basic Local Alignment Search Tool (megaBlast) to determine the region with the strongest alignment to other sequences in the BLAST database. The Hit Table output of BLAST lists all the matches to a particular sequence, in order of highest alignment. This output also identified the regions of alignment for each match. This region was identified in all sequences (Altschul 1990). Typically, this region fell between the forward and reverse primer within the sequence; however, at times the region fell on the primer, and therefore was another means by which we could recognize and remove flanking sequences that could skew final analyses. The program compared our unknown nucleotide sequences to known sequences in a database with over 61 million sequences, and calculated the statistical significance of matches (National Resource for Molecular Biology Information 2007). Following this final step, the portion of the sequence before and after the primers, restriction sites and the BLAST region were deleted. This left only the ~500 bp 16S rRNA insert for further analysis. An example of this step of editing is summarized in Figure 6.



▸NCBI/ BLAST/ blastn/ Formatting Results - TYFAYPYF01N [Reformat these Results] [Edit and Resubmit] [Sign in above to save your search strategy]

Job Title: 3 sequences (Ap10-2.L1.1.01.B01_07052311NQ...

```
# BLASTN 2.2.17 (Aug-26-2007)
# Query:  Ap10-2.L1.1.01.B01_07052311NQ    768     14     768    CEQ
# Database: nr
# Fields: query id, subject ids, % identity, alignment length, mismatches, gap opens, q. start, q. end, s. start, s. end, evalue, bit score
# 104 hits found
Ap10-2.L1.1.01.B01_07052311NQ   gi|82880242|gb|DQ285085.1|      92.94   524     14      21      134     647     1       511     0.0     741
Ap10-2.L1.1.01.B01_07052311NQ   gi|125967571|gb|EF392904.1|     92.12   533     18      24      134     657     1       518     0.0     732
```

```
>Ap10-2.L1.1.01.B01_07052311NQ    768     14    768   CEQ
            GGAGTTGTTCACACGGGCCAGTGAGCGCGCTAATACGATCTCACTATAGGGCGAATTGGAGCTCCCGCGTTG
CCACGCTACTAGAACTAGTGGATCCCCCGGGTCTTGCAGCACATTGTTGGAATTCGCCCTTAGAGTTTGATCCTGGCTC
AGAGTGAACGCTGGCGGCAGGCTAACACATGCAAGTCGAACGGCAGCACAGGGGAGCTTGCCTNGGGTGGCGAGTGG
CGGACGGGTGAGGAATACATGGGAATCTACCCTGTCGTGGGGGATAACGTAGGGAAACTTACGCTAATACCGCATACG
ACCGAGAGTTGAAAGCGGCGGACCGAAGGCGTCACGCGACTGGATGAGCCCATGTCGGATTAGCTAGTTGGCGGGGTA
AAGGCCCACCAAGGCGACGATCCGTAGCTGGTCTGAGAGGATGATCAGCCACACTTGGAACTGAGACACGGTCCAAGA
CTCCTACGGGAGGCAGCAGTGGGGGAATATTGGACAATGGGCGCAAGGTATCCCAGCCATGCCGCGTGGGTGAAAGA
AGGCCTTCGGGTTGTAAAGCCTTTTTGTCCCGGAAAGAAAAGCACGGGATTAAATACCCTCGTGTGATGACGGTACCCG
GAAGAAATACGCAACCGGCTACCTTTCGTGTCAAGCAGCCCCCGGTTCAAAAGGGCGAAAATCCCACAAGTTGGAATA
TTCAAGGCCTAATCGGATAACCGTCGACCCTCGAGCGCGCGGGCCCGGTTACCAAGCCTTTTTGTTTCCCTT
```

Figure 6: Editing Step 4
Identifying highest alignment region using megaBlast

```
>Ap10-2.L1.1.01.B01_07052311NQ    768     14    768   CEQ
AGTGAACGCTGGCGGCAGGCTAACACATGCAAGTCGAACGGCAGCACAGGGGAGCTTGCCTNGGGTGGCGAGTGGCG
GACGGGTGAGGAATACATGGGAATCTACCCTGTCGTGGGGGATAACGTAGGGAAACTTACGCTAATACCGCATACGAC
CGAGAGTTGAAAGCGGCGGACCGAAGGCGTCACGCGACTGGATGAGCCCATGTCGGATTAGCTAGTTGGCGGGGTAAA
GGCCCACCAAGGCGACGATCCGTAGCTGGTCTGAGAGGATGATCAGCCACACTTGGAACTGAGACACGGTCCAAGACT
CCTACGGGAGGCAGCAGTGGGGGAATATTGGACAATGGGCGCAAGGTATCCCAGCCATGCCGCGTGGGTGAAAGAAG
GCCTTCGGGTTGTAAAGCCTTTTTGTCCCGGAAAGAAAAGCACGGGATTAAATACCCTCGTGTGATGACGGTACCCGGA
AGAAATACGCAACCGGCTACCTTTCGT
>SSA12.1.18
GGGATGAACGCTAGCGGCAGGCTTAATACATGCAAGTCGTGGGGCAGCATGTCCCGCAGCAATGCGGGATGATGGCGA
CCGGCAAACGGGTGCGGAACACGTACACAACCTTCCTTTTAGTGGAGAATAGCCCAGGGAAACTTGGATTAATACTCC
GTAACATATAAGAAGTGGCATCACTTTTATATTAAAGCAGCAATGCGCTGGAAGATGGGTGTGCGGCTGATTAGATAG
TTGGCGGGGTAACGGCCCACCAAGTCGACGATCAGTAACTGGTGTGAGAGCACGACCAGTCACACGGGCACTGAGACA
CGGGCCCGACTCCTACGGGAGGCAGCAGTAAGGAATATTGGTCAATGGACGCAAGTCTGAACCAGCCATGCCGCG
>SSA12.1.23(RC)
AACGAACGCTGGCGGCGTGGATAAGACATGCAAGTCGAACGGACTAGGTTTGTAGCAATACAGGCTTGGTTCAGTGGC
GCACGGGTGCGTAACACGTGAGCAATCTGCCGGGAAGTGGGGGATAGCTCGCCGAAAGGCGAATTAATACCGCATATG
GCCATCCTTCGATTGGAGGAAAAGCCAAAGCAGCAATGCGCTTCCTGATGAGCTCGCGGCCTATCAGCTAGATGGCGG
GGTAACGGCCCACCATGGCGATGACGGGTAGCTGGTCTGAGAGGACGACCAGCCACACTGGAACTGAGATACGGTCCA
GACACCTACGGGTGGCAGCAGTCGAGAATTTTTCACAATGGGGGAAACCCTGATGGAGCGACGCCGCGTGGAGGATGA
AGGCCTTCGGGTTGTAAACTCCTGTCATGCGGGGAACAAGCAAGTGATAGTACCGCAAGAGGAAGAGACGGCTAACTC
TGTGCCAGCAGCCCC
```

Figure 7: Edited and Trimmed Sequences

The editing process outlined above was a crucial portion of this project. The sequences used for the DOTUR analysis, must have met all the criteria mentioned above. The software packages do not verify the input sequences provided to it. Therefore the software output provided must be validated by the editing process applied to the input. Figure 8 below is a flow chart that describes the procedures the raw sequences underwent and the various analyses performed.

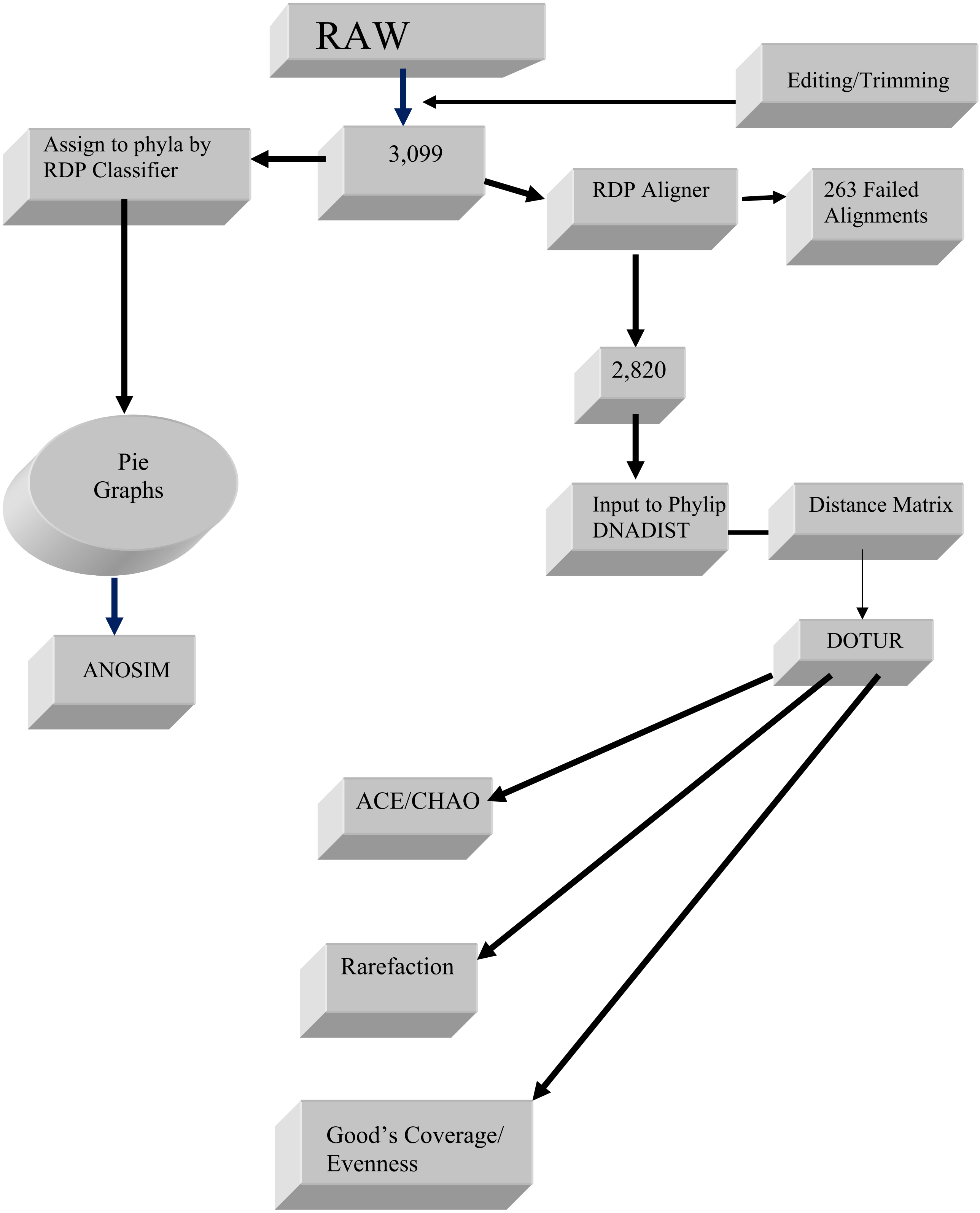


Figure 8: Schematic of Sequence Analysis

Analysis

The 3,099 sequences remaining after trimming and editing were aligned with the RDP release 9.57 aligner. This aligner was the only online program able to handle the capacity of sequences in this project. The data were separated into subsets representing the comparisons needed to answer the research questions. Data were sorted by control and planted mesocosms, by plant species, and by depth. These groupings of sequences were uploaded to the aligner, a process that took 10 days to complete.

The RDP Classifier program analysis was used to construct pie charts in Excel to address each research question. The pie charts divided the phyla represented in each community into 9 slices. At times, phyla with low representation were grouped together in order to make the graph more clear. Each of the pie charts also had a summary table for each phylum. The pie charts and tables are summarized in Chapter IV under their respective research questions. To verify that the community phyla classifications were statistically different, ANOSIM was performed on the RDP phylum classifications. If the p value was greater than .05, then the two communities being compared could not be statistically different.

The literature review presented the different parameters used in this project. The sequences remaining after trimming and editing were used to calculate richness parameters, evenness, and Good's coverage. However, these calculations become complicated with such a large number of sequences. DOTUR, the program used to calculate the parameters, required a distance matrix for execution. The aligned data was downloaded from RDP site in a Phylip format. The data subsets that numbered greater than 2,000 sequences were downloaded by the RDP staff due to program limitations.

This Phylip file for each subset of the data was used as an input file for the Phylip version 3.2 DNADIST program. This program used the Jukes-Cantor method to create a distance matrix. This distance matrix was used to run the DOTUR software.

Once the distance matrix was created, the file was saved as a distance file in the DOTUR program. This distance file was used to run the DOTUR program. 23 files of output data were created by DOTUR to include the ACE, CHAO 1, and rarefaction data. These files were used to create graphs and perform calculations to answer the research questions of this project.

DOTUR constructs *.c* files to plot collector's curves. These files are organized so that the first column is the number of sequences sampled. The next three columns for each evolutionary distance represented the mean parameter and the parameter's upper and lower 95% confidence interval bounds. At times, a confidence interval was difficult to define so a zero was placed in that particular spot (DOTUR 2005).

Each of the *.c* files for the parameters used in this project were used to construct collectors curves at the 3% evolutionary distance (species level), from other sequences within the samples, and the 20% evolutionary distance (phylum level), from other sequences within the samples. These graphs were used for comparison, and were able to address each of the research questions.

As previously mentioned, diversity consists of two parts: richness and evenness. The ACE, CHAO 1, and rarefaction data from DOTUR were used to construct curves to address richness. However, evenness was calculated by a simple formula. Evenness is considered the relative distribution of individuals among certain predefined units, such as species. There are numerous ways to determine evenness. This project used the most popular formula for evenness, the ratio of the Shannon index and the maximum value of observed OTUs when only one individual occupies each OTU (Kennedy 1995). The Shannon index was calculated by DOTUR. This was located in the Shannon *ltt* file. The average Shannon index for the 3% and 20% evolutionary distances were used in the evenness calculations. That value was divided by the LN(S), which is the total number of species at that evolutionary distance. The error was propagated by using the relative error from both the Shannon index and the S value. The 95% upper and lower confidence intervals were provided by DOTUR (Schloss & Handelsman 2005).

Good's coverage was determined by the traditional formula C= $1-\frac{n_1}{N}$ (Good 1953). N was defined as the community size and n_1 was defined as the number of phylotypes appearing only once, and C was Good's coverage. The coverage was calculated for each plant species, depth, control, compiled planted, and all the data.

Chapter IV: Results and Analysis

Overview

Data for all similar plant species were pooled to construct a 16S rRNA community for each comparison of interest: planted vs. unplanted, plant species, and depth within those groups. The main research objectives for this project were to determine if plant presence, plant species, or depth significantly impacted the makeup of the microbial community composition or diversity in the mesocosms. Several diversity parameters were used to answer these questions. This section summarizes the diversity parameters and analyses, and the outcomes of those analyses. This section begins with a general look at the diversity of the all of the sequence samples, and then is organized by research question.

The sequences fell into the categories summarized in Table 2, once all similar mesocosms were grouped together. The sequences were not evenly distributed due to the uneven planting scheme, wherein there were four columns with *S. atrovirens*, three with *E. erythropoda*, two with *C. comosa*, and three unplanted controls. The trimming and editing process, described in Chapter III, left 3,099 sequences. These sequences were assigned to phyla by the RDP Classifier program using an 80% match to sequences within the RDP database. Afterwards RDP alignment was executed, a total of 2,820 sequences were left for DOTUR analysis. 263 (8.5%) sequences failed to align due to RDP aligner program limitations (RDP staff 2007). Another 0.5% of the sequences fell into an Unclassified Root category, which is explained later in this section. Neither the sequences which failed to align nor the Unclassified Root sequences were used in the DOTUR analyses.

Table 2: Sequence Breakout

	Carex comosa	*Eleocharis erythropoda*	*Scirpus atrovirens*	Control	Total
Sequences after trimming and editing	506	756	1076	761	3099
Sequences after Alignment	471	695	959	695	2820

It was immediately evident that each microbial mesocosm community, even the control columns, was extremely species-rich in diversity, and that the sequences used to characterize this community came from just a small sample of the entire community. Table 3 below demonstrates that an average of 65% of all the sequences appeared only one time in each community at a sequence similarity of 97% (species level), and Table 4 shows an average of 25% appeared only one time at a sequence similarity of 80% (phylum level).

Table 3: Frequency Distribution of OTUs at 97% Similarity

			Number of OTUs with N_x sequences					
Community	Number of Sequences	Number of unique OTUs	N_1	N_2	N_3	N_4	N_5	$N_{>5}$
Scirpus atrovirens	959	657	566	56	12	8	6	9
Carex comosa	471	381	331	37	8	1	0	4
Eleocharis erythropoda	695	585	510	53	15	2	4	1
Control	695	528	442	62	13	4	2	5

Table 4: Frequency Distribution of OTUs at 80% Similarity

			Number of OTUs with N_x sequences					
Community	Number of Sequences	Number of unique OTUs	N_1	N_2	N_3	N_4	N_5	$N_{>5}$
Scirpus atrovirens	959	197	68	40	29	17	10	33
Carex comosa	471	130	51	30	12	10	4	23
Eleocharis erythropoda	695	190	77	35	23	16	10	29
Control	695	178	71	39	15	13	12	28

The first step was to characterize the community composition. This was performed by comparing the sample sequences to the RDP database of known sequences. Figure 9 depicts the various phyla the 3,099 sequences fell into using RDP Classifier program. This figure illustrates the community composition of a summation of all the sequences. This summation of microbial community composition across the mesocosms models the soil of the constructed wetland at WPAFB.

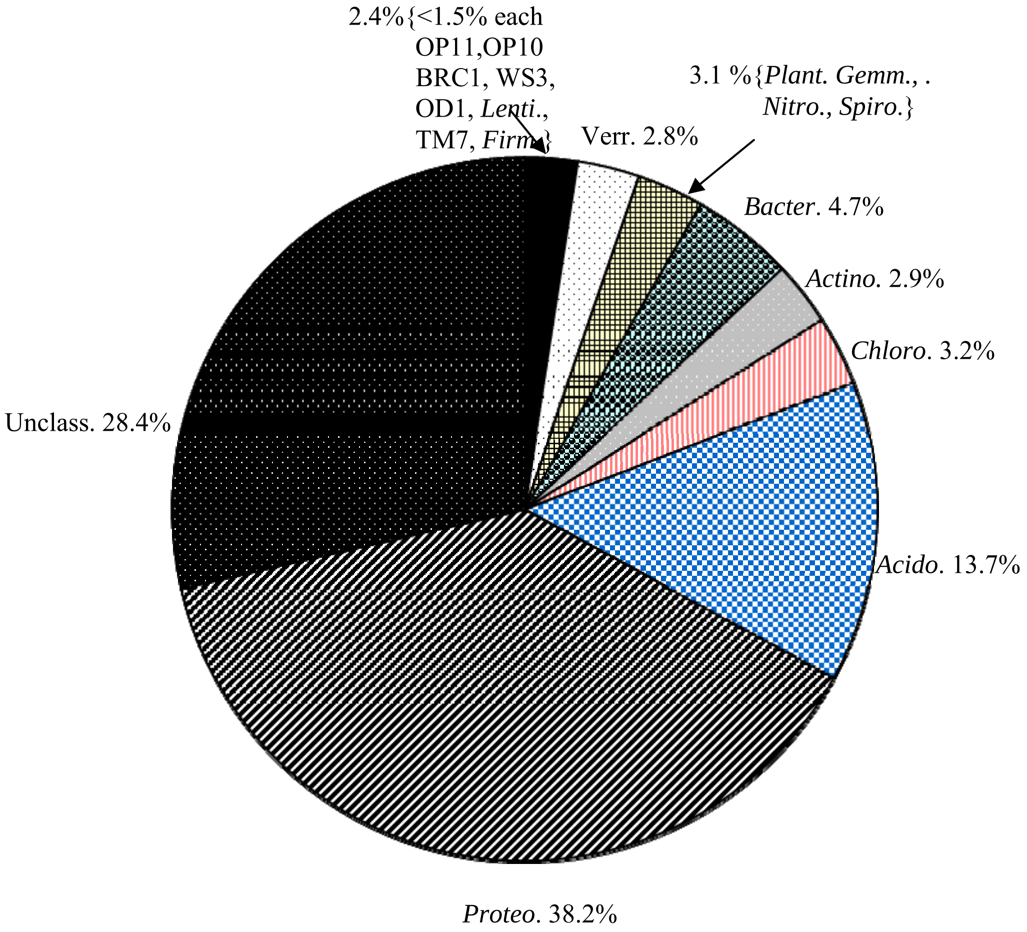


Figure 9: Phyla Classification for all Data using RDP Classifier
Abbreviations: *Acido., Acidobacteria; Actino., Actinobacteria; Bacter., Bacteroidetes; Chloro., Chloroflexi; Firm., Firmicutes; Gemma., Gemmatimonadetes; Lenti., Lentisphaerae; Nitro., Nitrospira; Plant., Planctomycetes; Proteo., Proteobacteria; Spiro., Spirochaetes;* Unclass., Unclassified Bacteria*; Verr., Verrucomicrobia.*

Of the 3,099 sequences used in the RDP classifier analysis, 99.48% were identified as belonging to the domain Bacteria with 18 different distinct phyla and an Unclassified Bacteria category. The remaining 0.52% fell into an Unclassified Root category. Unclassified Root refers to sequences for which the RDP Classifier Program could not determine whether they were bacterial16S rRNA. These may have been non-16S genes or rRNA genes from non bacteria or sequences of low quality (RDP Staff

2007). This category was not shown on any of the pie charts in this section and was also eliminated from DOTUR analyses.

28.4% of the sequences fell into the Unclassified Bacteria category, which meant that random subsets of the query sequence did not match sequences within the RDP database greater than or equal to 80% of the time. The remaining sequences were assigned to a phylum. The largest group, 38.2%, was *Proteobacteria.* Although phylum richness was high with 19 different phyla represented, the abundance was not even. The prevalent phyla represented, other than the *Proteobacteria,* were *Acidobacteria*, 13.7%, and *Bacteroidetes*, 4.7%. It is important to mention that phyla known to contain dehalogenators, *Chloroflexi* and *Firmicutes*, were present in very small numbers.

The second step was to characterize the diversity of the sample sequences. This analysis was performed using DOTUR, where the sample sequences were compared to each other. A rarefaction curve, the ACE, and Chao 1 parameter, were calculated for the entire data set. The figures for the species and phylum levels are below. The species graph did not reach an asymptote; however, the phylum level graph did reach an asymptote for the ACE and Chao 1 estimators and the rarefaction curves. The lack of an asymptote indicates high richness and that the total population was undersampled. It was apparent the total community was very diverse, and that the community as a whole was probably undersampled in this project, especially at the species level.

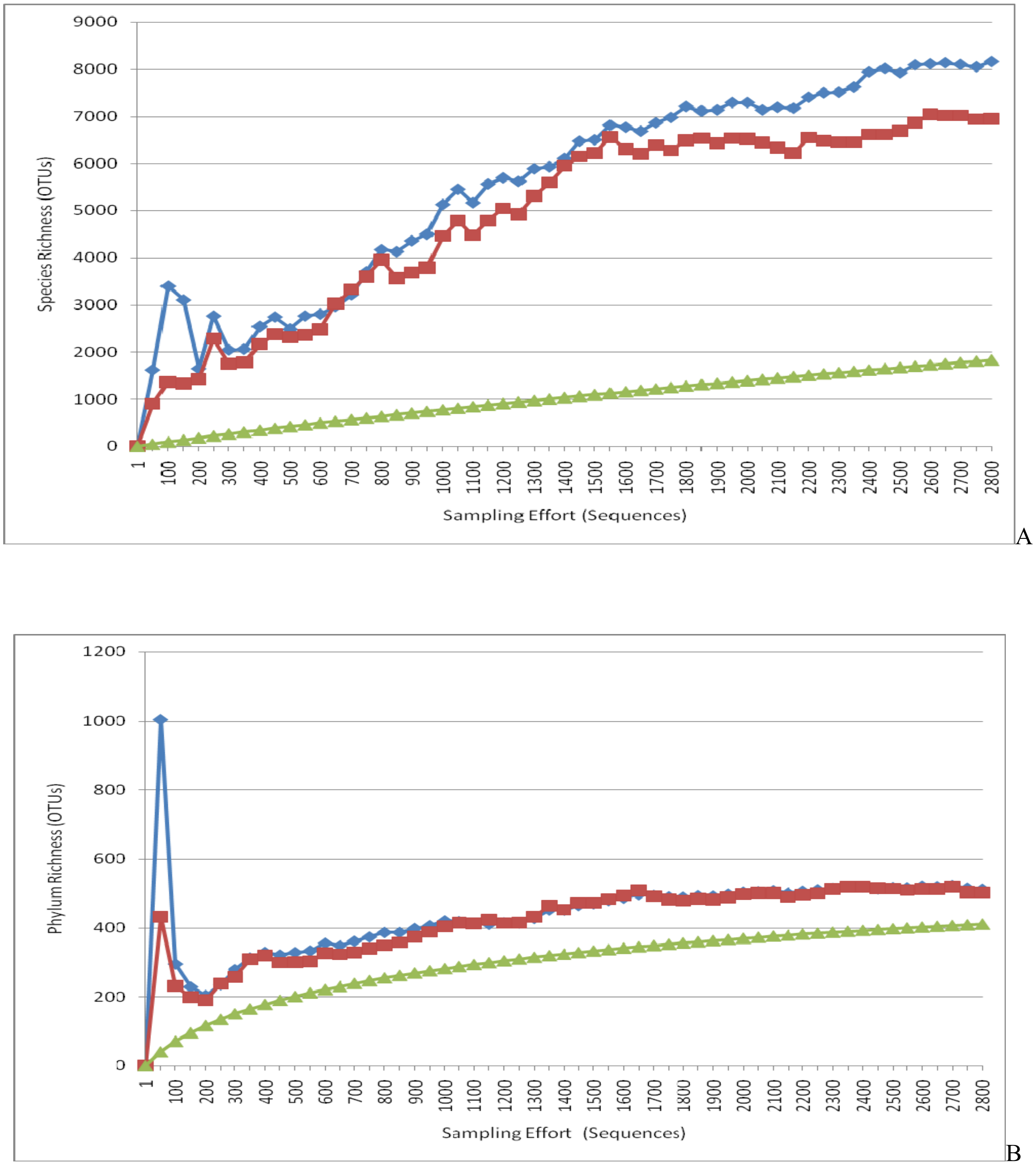


Figure 10: All Richness Estimates and Rarefaction Curves
Ace (diamonds) and Chao (square) richness estimators at the species level (A) and phylum level (B) for all the data. Rarefaction values (triangles) based on observed OTUs.

Another important point to establish was that the sample effort was adequate to provide quality data for interpretation. Good's coverage was calculated for each comparison. Figure 11, below, summarizes the coverage for the entire data set. As expected, the phylum-level coverage was high relative to the species coverage. The phylum coverage averaged 92%; therefore, the parameters calculated for the phylum level come from a population that had been sampled at a high level.

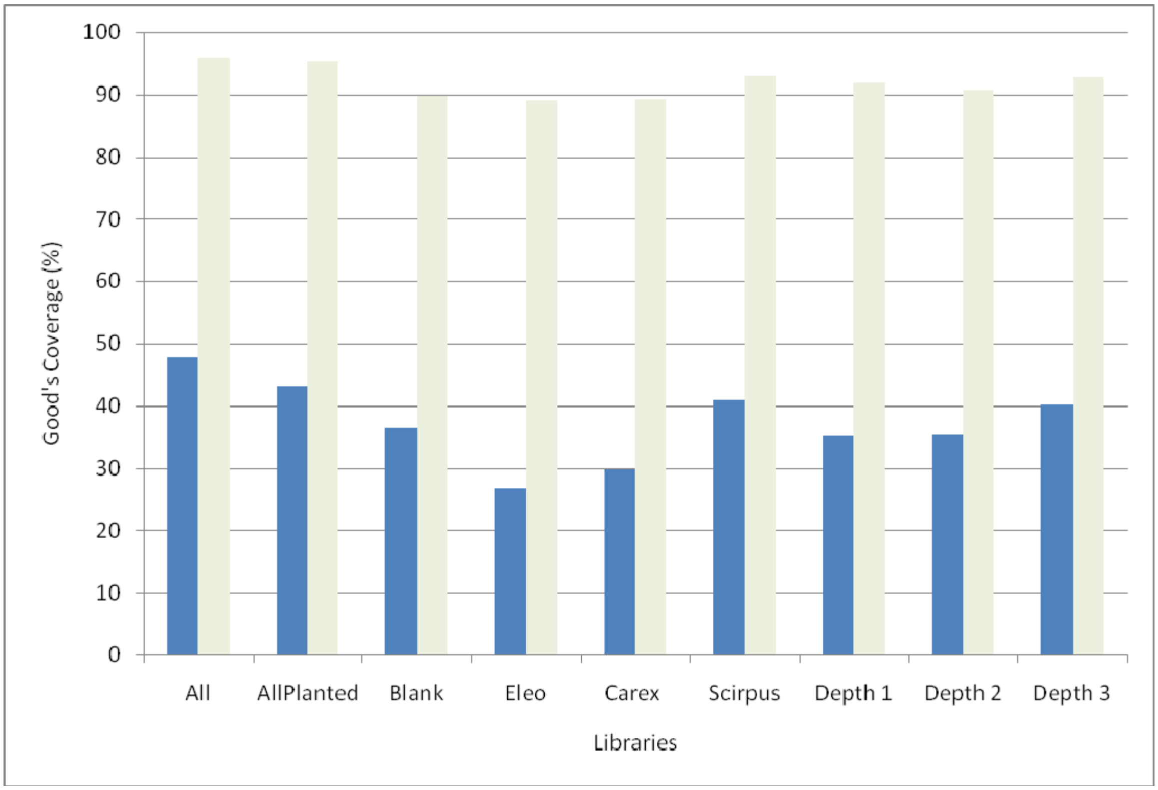


Figure 11: Good's Coverage
Light green bars represent the phylum level. Blue bars represent the species level.

Evenness was also an important aspect that was investigated. Figure 12 below summarizes the results for the entire data set. Evenness was calculated with the Pielou equation presented in Chapter III. The error bars represent the propagated error for each constituent in the formula. The error was calculated by DOTUR.

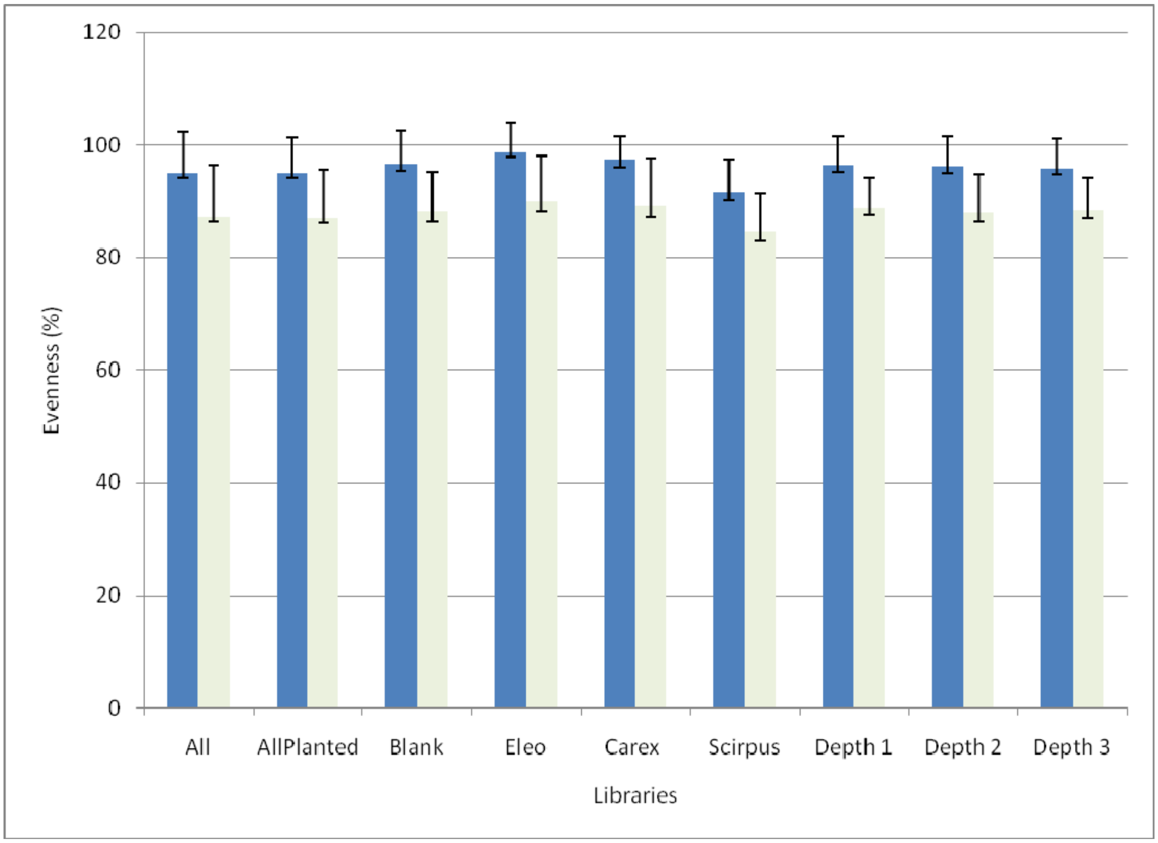


Figure 12: Evenness
Species level (Blue) and the phylum level (Light green)

<u>Research Objective 1</u>: Determine the effects of plant presence with regards to microbial diversity and dominance

The first step was using the RDP classifier function to identify all the DNA sequences that could be matched to a known species of microorganism within the RDP database. This characterized the community composition for both communities. The results are summarized in Figure 13 below.

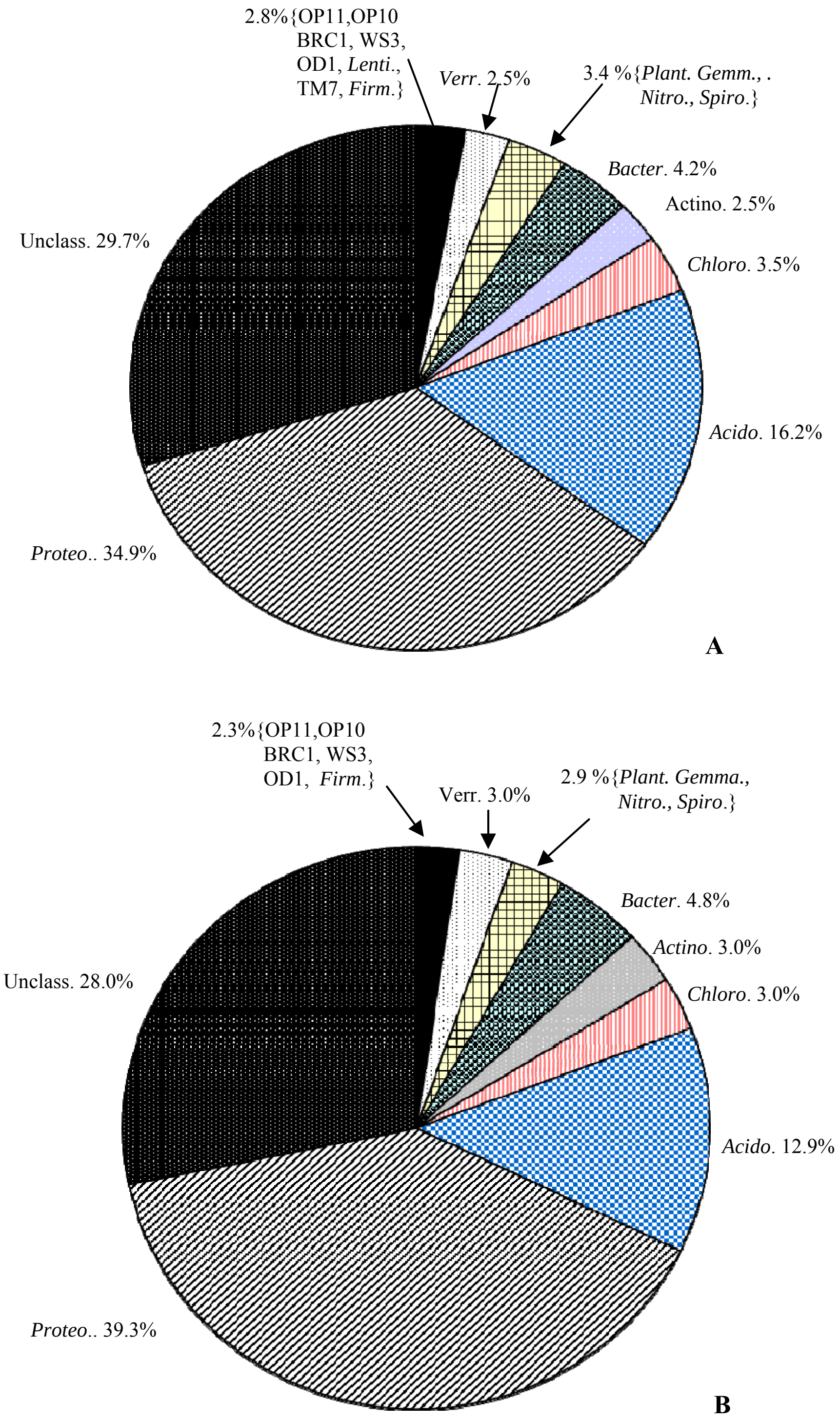


Figure 13: Phyla Classification for all Control sequences (A) and all Planted sequences (B) using RDP Classifier.

Abbreviations: *Acidobacteria; Actino., Actinobacteria; Bacter., Bacteroidetes; Chloro., Chloroflexi; Firm., Firmicutes; Gemma., Gemmatimonadetes; Lenti., Lentisphaerae; Nitro., Nitrospira; Plant., Planctomycetes; Proteo., Proteobacteria; Spiro., Spirochaetes*; Unclass., Unclassified Bacteria; *Verr., Verrucomicrobia.*

The charts illustrate that the microbial composition for the known sequence matches for both the planted and control data are very similar even though the planted community had four times the sequences as the control. Table 5 below summarizes the actual percentage of each phylum. There are several interesting trends that can be noticed from the table. Two phyla were represented in the control community, but were not found in the planted community. The phyla TM7 and *Lentisphaerae* each appear one time. Since the sequences produced during this experiment are representative of the dominant phyla within the soil samples, the presence of one individual was important to document.

Table 5: Phyla Classification Percentages (Control vs. Planted)

Phyla	Control	Planted
TM7	0.13	0
OP11,OP10,OD1,WS3,BRC1	1.18	0.98
Verrucomicrobia	2.5	2.95
Firmicutes	1.31	1.32
Spirochaetes	0.26	0.38
Plantomycetes	0.92	0.86
Bacteroidetes	4.2	4.79
Lentisphaerae	0.13	0
Actinobacteria	2.5	3.04
Nitrospira	1.18	1.07
Chloroflexi	3.55	3.04
Acidobacteria	16.16	12.87
Proteobacteria	34.95	39.35
Gemmatimonadetes	1.05	0.64
Unclassified Bacteria	29.7	28.06

In order to understand whether microbial community composition differed statistically, we analyzed the RDP Classifier data using ANOSIM. Analysis revealed no significant differences between the planted and control data, (n=5000 permutations; p=0.75). The outcome was the same when unclassified sequences were dropped from the phylum level analysis.

The second step was to characterize diversity using DOTUR analysis. Evenness was summarized in Figure 12. There was high evenness for both communities for the phylum and species level. This combined with the fact that the Good's coverage at the species level was low, indicated that the species level was vastly undersampled. However, the phylum level Good's coverage was high which indicated that the sampling effort was adequate enough to make a confident assessment at this level. Richness parameters used for analysis in this project had some differences between the planted and control communities. Figure 14 shows estimates of richness at the 3% distance level for species (A) and 20% distance level for phylum(B).

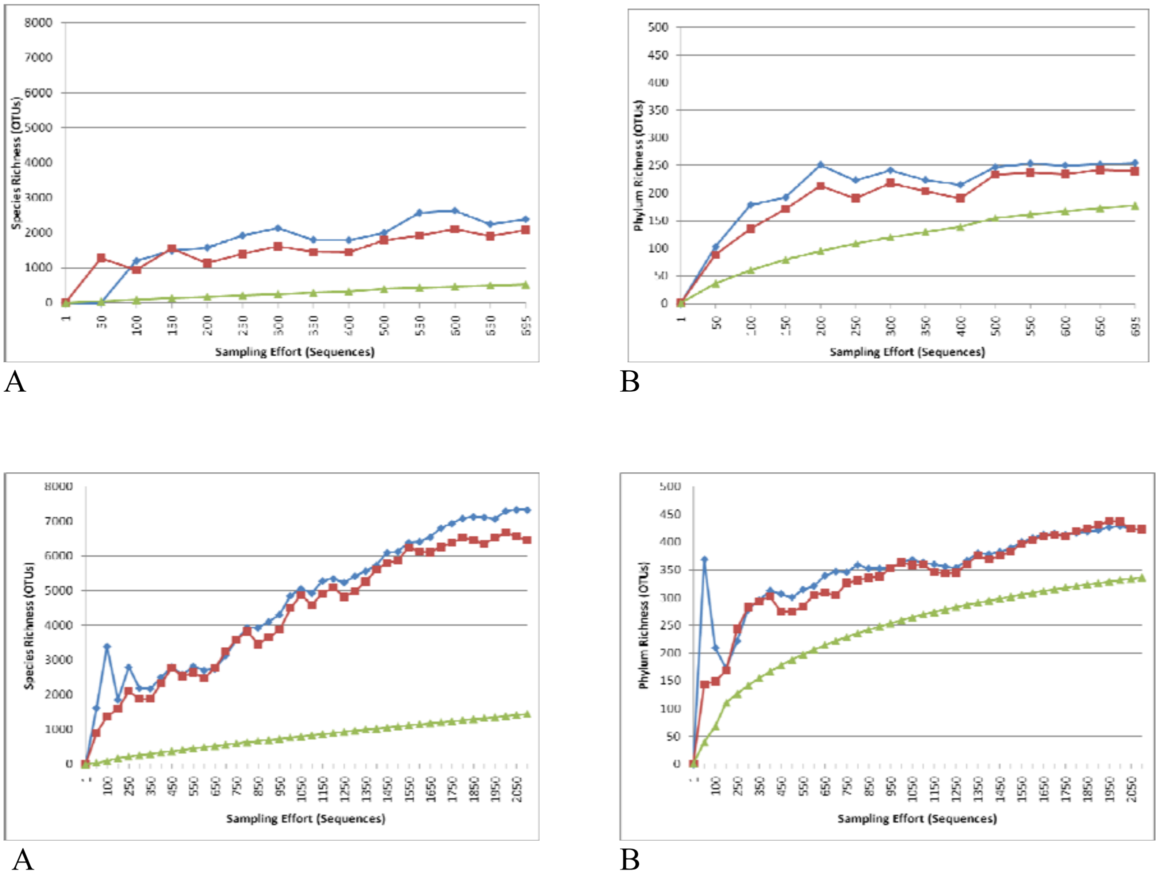


Figure 14: Control and Planted Data Richness Estimates and Rarefaction Curves
Chao (diamonds) and ACE (squares) richness estimators at the species (A) and the phylum level (B) for Control (top) and Planted (bottom) data. Rarefaction values (triangles) based on observed OTUs.

From these graphs we can make some important observations. The planted sequences had a much higher richness estimate than the controls. The Chao 1 and ACE estimators were 4500 units higher in the planted sequences at the species level. However, the species level graphs for all three richness parameters never reached an asymptote. This again shows us that the species level was undersampled. In the phylum graphs the Chao 1 and ACE estimators were somewhat closer for the planted and control communities, and both the estimators and the rarefaction curve did asymptote. The planted community was still much higher than the control; however, this could be due the sample size for the planted data, 1430 sequences higher than the control. Therefore, a look at the rarefaction curves for 695 random sequences for each group was warranted.

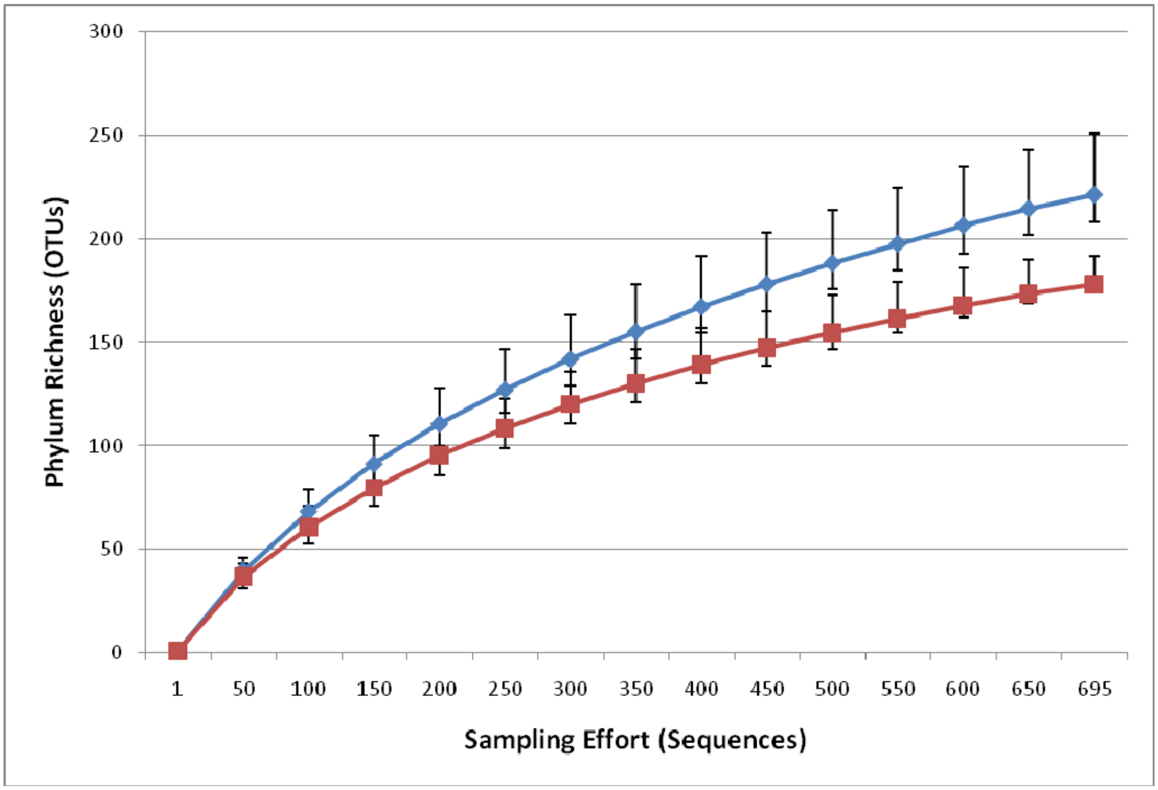


Figure 15: Phylum Level Rarefaction Curve for Control and Planted Data
Planted (diamonds) and Control (squares) rarefaction values based on observed OTUs at the phylum level.

In the phylum level analysis, the rarefaction curve does approach an asymptote for both data sets. This indicates that the sampling effort was adequate to make a clear and good estimate of richness at the phylum level. The planted sequences had a higher richness than the control data, even when a random 695 sequences were taken for both the planted and control communities. Also the error bars here show that at the lower sample size of less than 350 the communities are not statistically different because they overlap. But as the sample size increase above 350 sequences, the error bars do not overlap and the richness values are statistically different. This analysis clearly shows that while the microbial community composition of known microorganisms at the phylum level did not change for the planted versus the control libraries, the richness was affected by plant presence at the phylum level.

Another trend seen here was that the confidence intervals did not get smaller as the sample size increased, as expected from typical statistic trends. This indicates that with increased sample size the variance of the data also increases. This phenomenon indicates that the communities are extremely rich, so that a true estimate of variance can never be made.

Research Objective 2: Determine the effects of plant species with regards to microbial diversity and community composition

The first step was using the RDP classifier function to identify all the DNA sequences that could be matched to a known species of microorganism within the RDP database. This characterized the community composition for the plant species communities. The results are summarized in Figure 16 below.

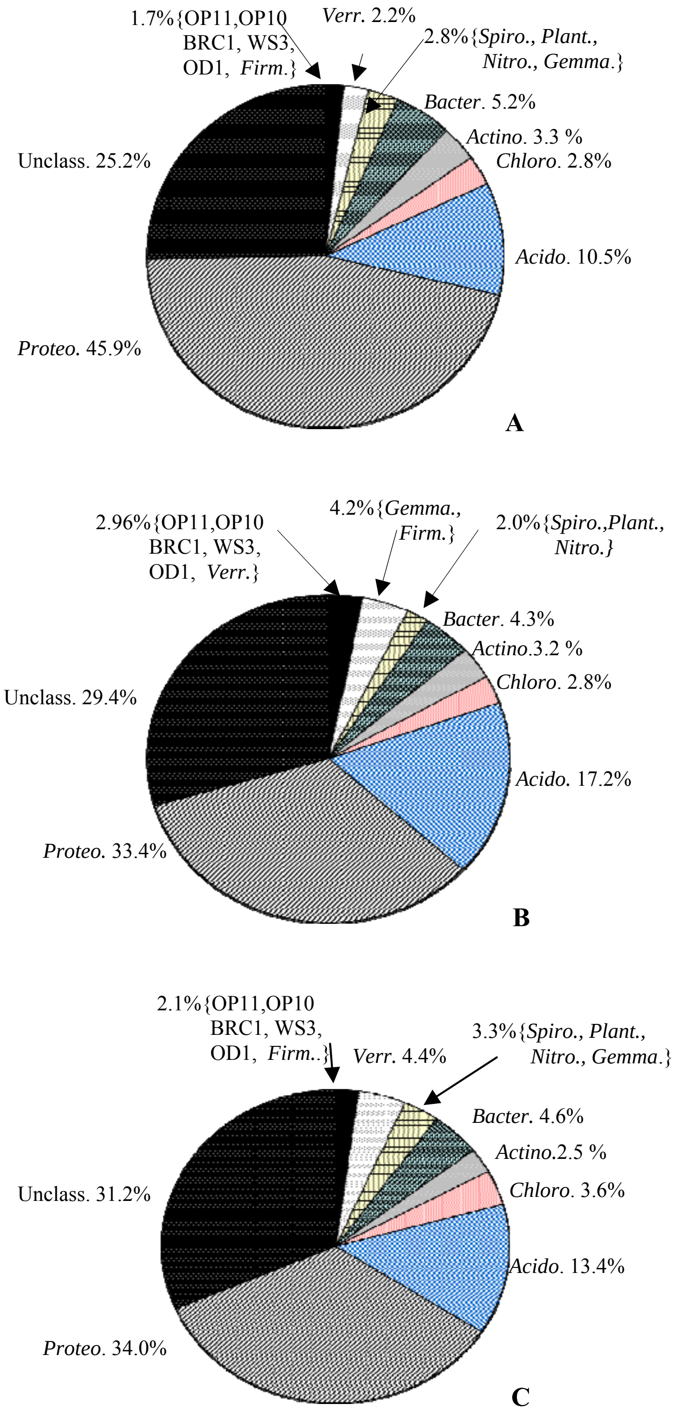


Figure 16: Phyla Classification for *all Scirpus atrovirens* sequences (A), all *Carex comosa* sequences (B), and all *Eleocharis erthyropoda* sequences (C) using RDP Classifier Abbreviations: *Acidobacteria; Actino., Actinobacteria; Bacter., Bacteroidetes; Chloro., Chloroflexi; Firm., Firmicutes; Gemma., Gemmatimonadetes; Lenti., Lentisphaerae; Nitro., Nitrospira; Plant., Planctomycetes; Proteo., Proteobacteria; Spiro., Spirochaetes*; Unclass., Unclassified Bacteria; *Verr., Verrucomicrobia.*

Table 6: Phyla Classification Percentages for *Scirpus atrovirens* sequences (A), *Carex comosa* sequences (B), and *Eleocharis erythropoda* sequences (C)

Phyla	***Carex***	***Eleocharis***	***Scirpus***
TM7	0	0	0
OP11,OP10,OD1,WS3,BRC1	0.59	0.92	1.21
Verrucomicrobia	2.37	4.36	2.23
Firmicutes	3.36	1.19	0.46
Spirochaetes	0.2	0.53	0.37
Planctomycetes	0.4	0.93	1.02
Bacteroidetes	4.35	4.63	5.2
Lentisphaerae	0	0	0
Actinobacteria	3.16	2.51	3.34
Nitrospira	1.38	1.32	0.74
Chloroflexi	2.77	3.57	2.79
Acidobacteria	17.19	13.36	10.5
Proteobacteria	33.4	33.99	45.91
Gemmatimonadetes	0.79	0.53	0.65
Unclassified Bacteria	29.44	31.22	25.18

The purpose of this analysis was to note any changes in microbial composition between the different species of plants at the phylum level. Although the composition was very similar, there were some slight differences. The phylum *Firmicutes* represents 3.4% of the sequences of the *Carex comosa* mesocosm samples, but only 0.46% and 1.2% of the *Scirpus atrovirens* and *Eleocharis erythropoda* communities, respectively. *Firmicutes* is a phylum known to contain dehalogenators. Since this mesocosm study mimics a constructed wetland treating a PCE and TCE plume, the presence of dehalogenators was expected.

The only other differences were with the phyla *Verrucomicrobia* and *Proteobacteria*. The *Eleocharis erythropoda* mesocosm samples had a 4.4% representation of *Verrucomicrobia* while the *Carex comosa* and *Scirpus atrovirens* samples had 2.4% and 2.2% respectively. The most prevalent phylum in all the plant species was *Proteobacteria.* However, *Scirpus atrovirens* had 45.9% representation while the other two species of plant had only an average of 33.7% representation.

In order to understand whether microbial community composition differed statistically, we analyzed the RDP Classifier data using ANOSIM. Analysis revealed no significant differences between the plant species data, (n=5000 permutations; p=0.21). The outcome was the same when unclassified sequences were dropped from the analysis.

The second step was to characterize diversity using DOTUR analysis. Evenness was summarized in Figure 12. There was high evenness for all the communities for the phylum and species level. This combined with the fact that the Good's coverage at the species level was low, indicates that the species level was vastly undersampled. However, the phylum level Good's coverage was high which indicated that the sampling effort was adequate enough to make a confident assessment at this level. Richness parameters used for analysis in this project had some differences between the plant species communities. Figure 17 shows estimates of richness at the 3% distance level for species (A) and 20% distance level for phylum(B).

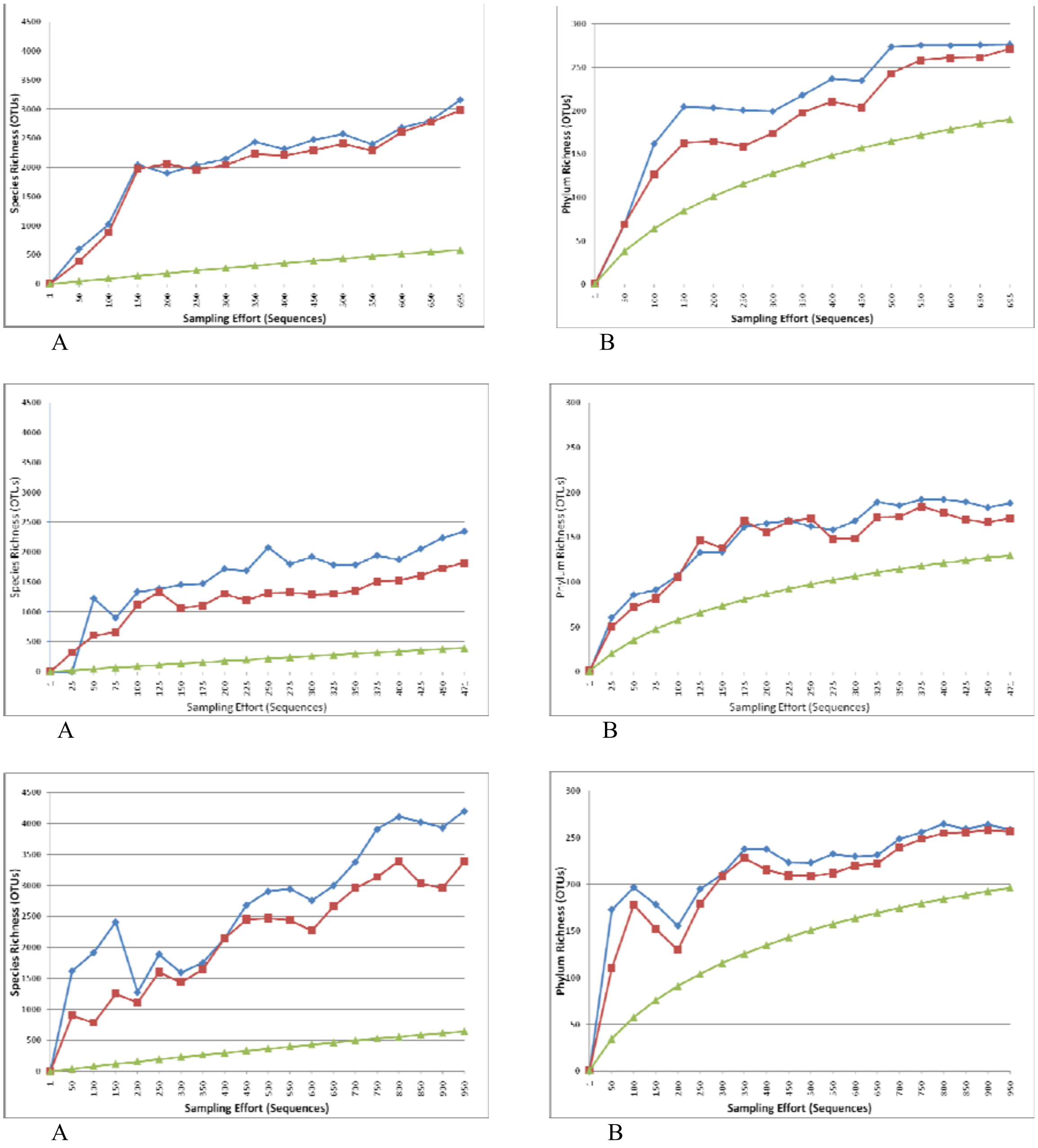


Figure 17: Plant Species Data Richness Estimates and Rarefaction Curves Chao (squares) and ACE (diamonds) richness estimators at the species (A) and the phylum level (B) for *Eleocharis erythropoda* (top), *Carex comosa* (middle), and *Scirpus atrovirens* (bottom) data. Rarefaction values (triangles) based on observed OTUs.

The richness estimators showed some interesting trends. The ACE estimator predicted the highest richness in all cases, while the observed richness (as show by the rarefaction curves) was always well below either estimator. This is because the rarefaction curve illustrated the real richness present in the samples. The ACE and CHAO 1 estimators estimate the true richness in the community that was sampled. The *Scirpus atrovirens* community had a much higher ACE and CHAO 1 estimate than the other two communities. This shows that more OTUs were identified in this community. These richness estimators had a slight difference in their values, suggesting plant species had an effect on microbial richness in the mesocosms. *Eleocharis erythropoda* had the second highest richness, while *Carex comosa* had the lowest richness of the species of plant. However, the estimators did vary with sampling effort and were not vastly different from each other. This was expected because all the species of plant used in this project were from the same family and had the same growth habit.

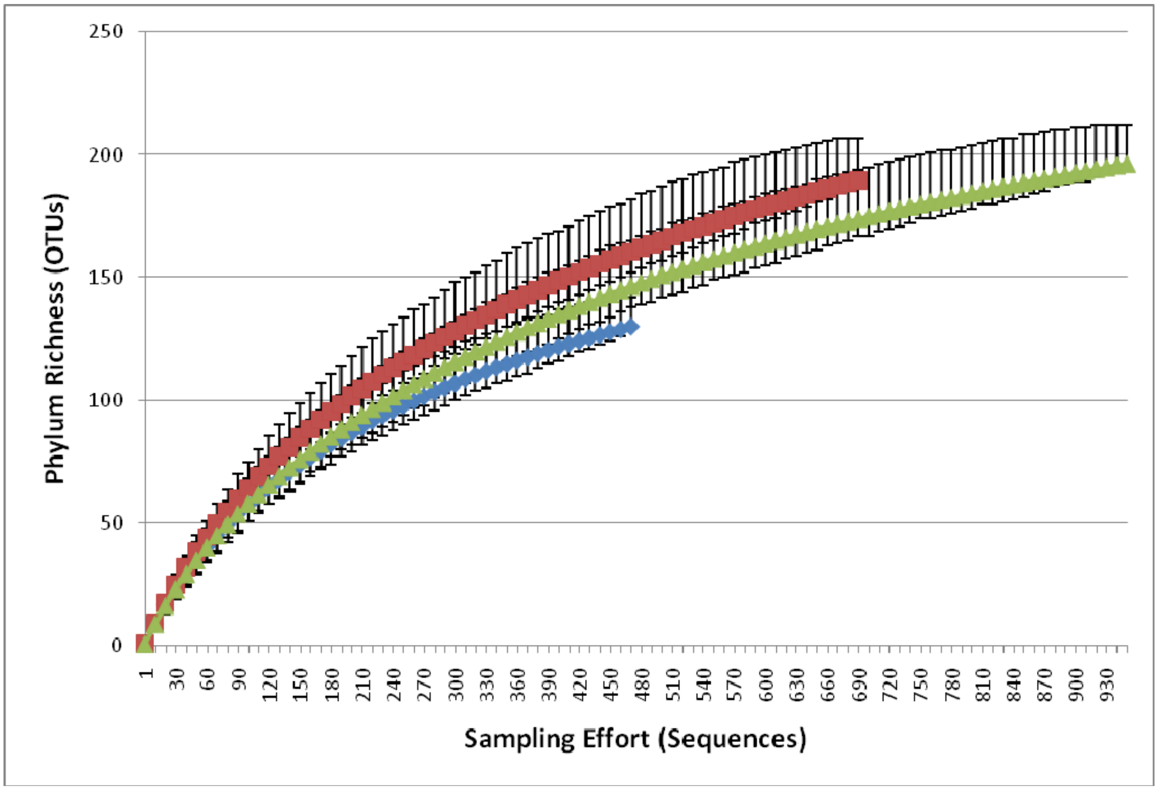


Figure 18: Phylum Level Plant Species Data Rarefaction Curve
Rarefaction values based on observed OTUs. *Eleocharis erythropoda* (squares) and *Carex comosa* (diamonds) *Scirpus atrovirens* (triangles) at phylum level.

In Figure 17, it was important to notice that the species level rarefaction data never reached an asymptote, indicating undersampling of the total population. However, the phylum level rarefaction data did reach an asymptote for each of the plant species. Figure 18 summarized the phylum rarefaction data calculated by DOTUR from the samples taken. The *Eleocharis erythropoda* data has the highest richness followed by *Scirpus atrovirens*. *Carex comosa* had the lowest richness. The error bars on this figure represent the 95% confidence interval. The error bars for all three plant species overlap, except the *Eleocharis* and *Carex* communities. Therefore, the *Eleocharis* and *Carex* communities have a difference in phylum richness. The communities are not sampled evenly but the trend, illustrated in Figure 18, seems to be that there was less overlap as the sample size increased. This could indicate that the plant species do have a richness difference at higher sample sizes.

The unexpected trend of stable confidence intervals with increasing sample size, previously discussed in Research Objective 2, was also seen here. This indicates that the true richness of these communities is extremely high. The sample size used here was not sufficiently large to establish a consistent estimate of variance.

Research Objective 3: Determine the effects of soil depth with regards to microbial diversity and community composition

The first step was using the RDP classifier function to identify all the DNA sequences that could be matched to a known species of microorganism within the RDP database. This characterized the community composition for the depth communities. The results are summarized in Figure 19 below.

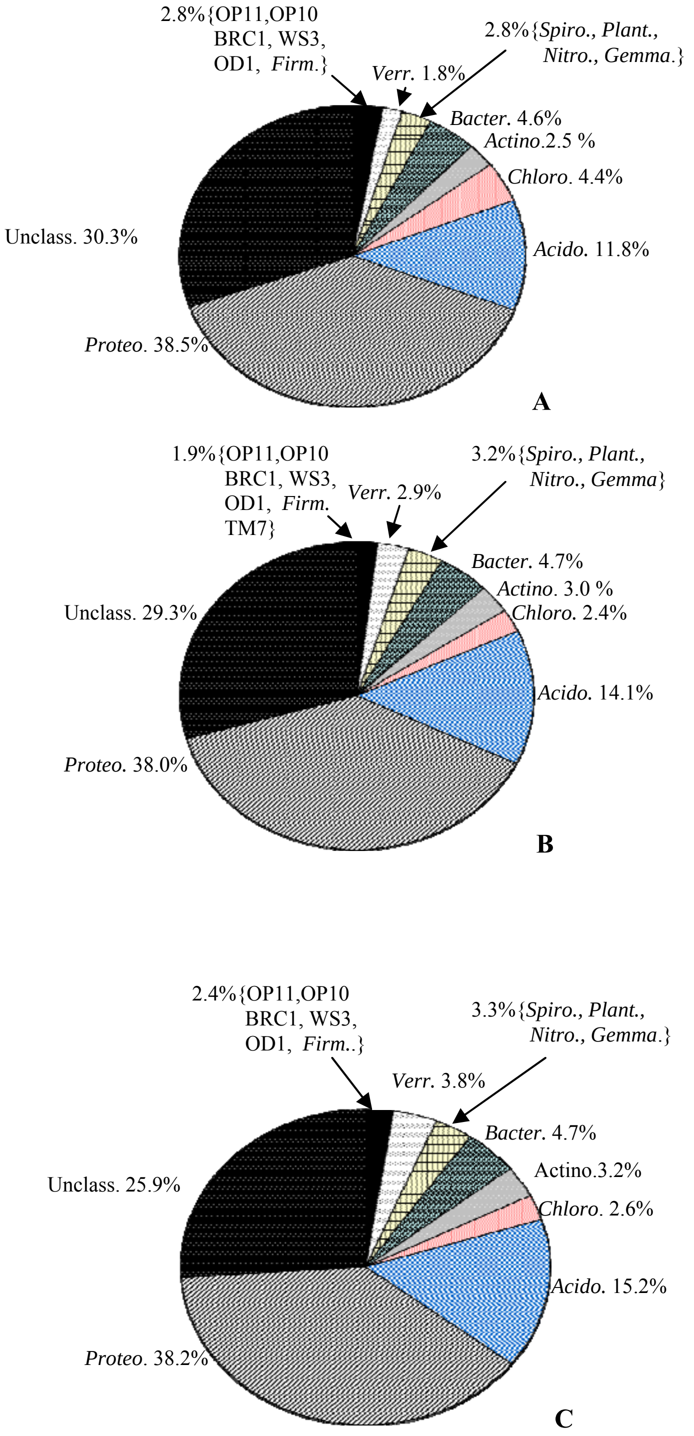


Figure 19: Phyla Classification for *all* Depth 1 sequences (A), all Depth 2 sequences (B), and all Depth 3 sequences (C) using RDP Classifier
Abbreviations: *Acidobacteria; Actino., Actinobacteria; Bacter., Bacteroidetes; Chloro., Chloroflexi; Firm., Firmicutes; Gemma., Gemmatimonadetes; Lenti., Lentisphaerae; Nitro., Nitrospira; Plant., Planctomycetes; Proteo., Proteobacteria; Spiro., Spirochaetes*; Unclass., Unclassified Bacteria; *Verr., Verrucomicrobia.*

Table 7: Phyla Classification Percentages

Phyla	**Depth 1**	**Depth 2**	**Depth 3**
TM7	0	0.11	0
OP11,OP10,OD1,WS3,BRC1	1.09	1.14	0.98
Verrucomicrobia	1.81	2.86	3.8
Firmicutes	1.72	0.69	1.43
Spirochaetes	0.45	0.23	0.36
Planctomycetes	0.90	0.92	0.80
Bacteroidetes	4.62	4.69	4.73
Lentisphaerae	0	0.12	0
Actinobacteria	2.54	2.97	3.21
Nitrospira	0.90	1.49	1.07
Chloroflexi	4.35	2.40	2.59
Acidobacteria	11.78	14.07	15.25
Proteobacteria	38.50	37.99	38.27
Gemmatimonadetes	0.54	0.57	1.07
Unclassified Bacteria	30.34	29.29	25.96

The purpose of this analysis was to investigate whether there were differences between microbial community compositions between the different depths. Although the composition was very similar there were some slight differences. Depth 1 correlates to the bottom of the mesocosm. The *Chloroflexi* population represents 4.4% of the Depth 1 samples taken. The middle and top depth, Depth 2 and Depth 3 respectively, were both around 2.5%. This could indicate that the bottom layers of the mesocosms are richer in *Chloroflexi*. It is also important to mention that the prevalent phylum in all the depths was *Proteobacteria*, an average of 38.2%, and the Unclassified Bacteria made up an average of 28.5% in all the depth communities.

In order to understand whether there were statistically significant differences in microbial community composition, we analyzed the RDP Classifier data using ANOSIM. Analysis revealed no significant differences among depths, (n=5000 permutations;

p=0.31). The outcome was the same when unclassified sequences were dropped from the analysis.

The second step was to characterize diversity using DOTUR analysis. Evenness was summarized in Figure 12. There was high evenness for all three communities for the phylum and species level. This combined with the fact that the Good's coverage at the species level was low, indicates that the species level was vastly undersampled. However, the phylum level Good's coverage was high which indicates that the sampling effort was adequate enough to make a confident assessment at this level. Richness parameters used for analysis in this project had some differences between the plant species communities. Figure 20 shows estimates of richness at the 3% distance level for species (A) and 20% distance level for phylum(B).

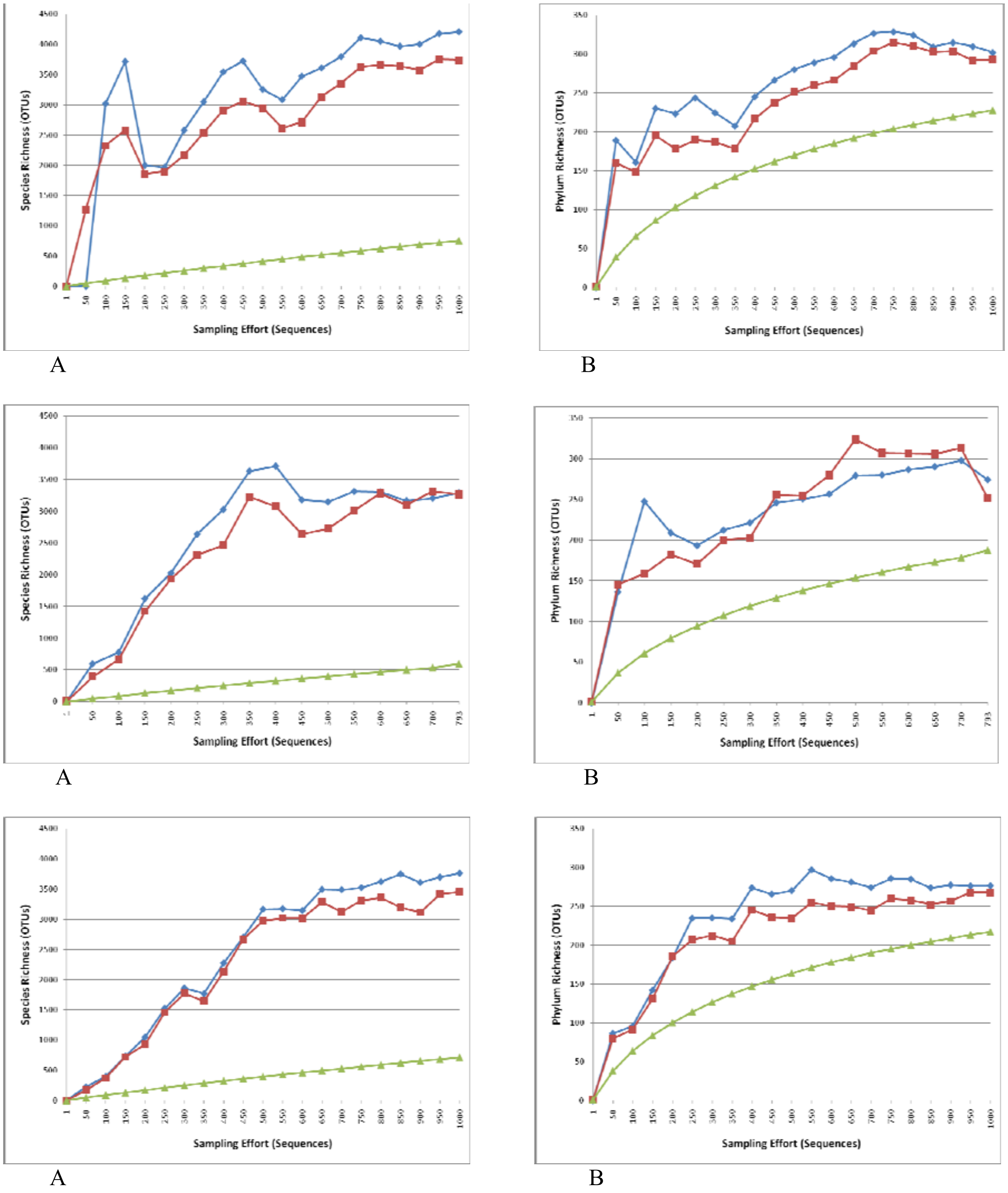


Figure 20: Depth Data Richness Estimates and Rarefaction Curves
Chao (squares) and ACE (diamonds) richness estimators at the species (A) and the phylum level (B) for Depth 1 (top), Depth 2 (middle), and Depth 3 (bottom) data. Rarefaction values (triangles) based on observed OTUs.

The richness estimators summarized in these graphs showed some slight trends but no strong evidence that the depths were different in diversity. The middle depth was slightly lower in both the ACE and Chao 1 estimators. This indicates that the middle depth had lower species richness than the top and bottom layers. In Figure 20, it was important to notice that the species level rarefaction curve never reached an asymptote indicating undersampling of the total population. However, the phylum level rarefaction data did reach an asymptote for each of the plant species. A closer look at the phylum level rarefaction data below, in Figure 21, uncovers an interesting trend.

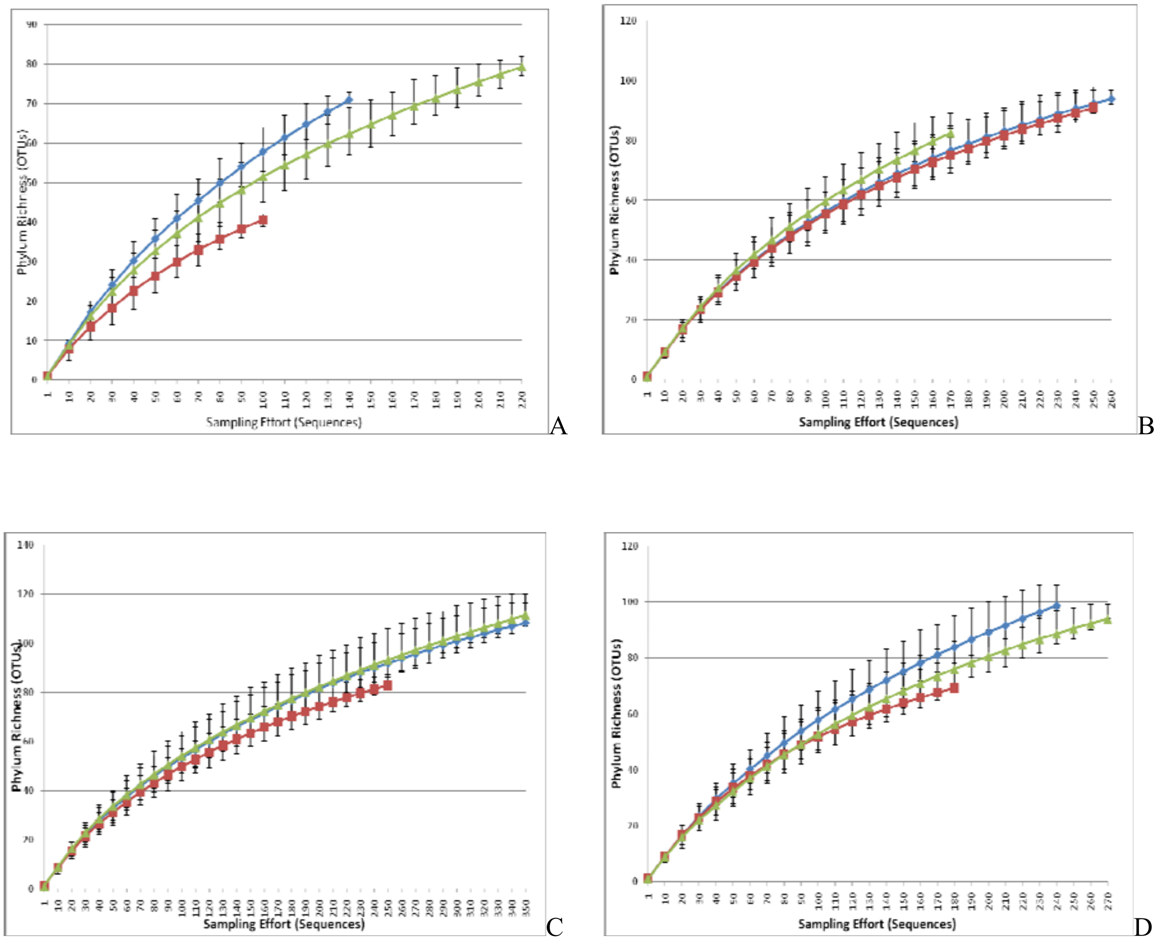


Figure 21: Community Phylum Level Depth Rarefaction Curves
Rarefaction values based on observed OTUs at phylum level. Middle depth (squares), Bottom depth (diamonds), and Top depth (triangles). *Carex comosa* (A); *Eleocharis erythropoda* (B); *Scirpus atrovirens* (C); Control (D).

The middle depth in all libraries reached an asymptote at a lower value. This indicates that the middle depth had lower diversity than both the top and bottom layers at the phylum level. The error bars on these curves represent the 95% confidence interval calculated by DOTUR. The error bars all overlap more than 50% except in the *Carex* and control communities. This indicates that, for these two communities, the middle layer was significantly different in richness than the other two layers. However, as sampling effort increased the layers in all communities did start to split apart. This trend indicates that the middle layer of all the communities was lower and this trend should be investigated in future research. Also the trend previously mentioned in the first two objectives of stable confidence intervals also applied. The intervals did not get smaller with increased sampling effort. This indicated that the total population was extremely diverse and a much larger sample size would have to be taken. All richness estimator and rarefaction curves are included in Appendix G.

Chapter V: Conclusions and Recommendations

Overview

This chapter summarizes the conclusions and recommendations from this research. All three research objectives are reviewed and the conclusions for each are discussed. Also this chapter reviews the significance of this research and the contribution it made to the literature in this area. This chapter ends with recommendations for further research.

This research focused on characterizing the microbial community composition and diversity for soil communities in constructed mesocosms prior to contamination of PCE. The mesocosm construction was based on a subsurface flow wetland remediating a PCE and TCE plume on WPAFB, OH, but the mesocosms were built with uncontaminated soil. Evidence had already shown that the wetland was remediating the groundwater plume (Amon 2007). Therefore, it was expected that phyla containing known dehalogenators would be represented in the non-contaminated sample sequences. Dehalogenators and other anaerobic organisms facilitate the first stage of PCE and TCE remediation.

From the 3,099 sample sequences used for RDP phyla classification, 3.33% of the sequences belonged to two phyla known to contain dehalogenators and anaerobic bacteria. The phylum *Chloroflexi* contains an organism, Dehaloccoides, that is a known dehalogenator, and the phylum *Firmicutes* contains anaerobic organisms with low G+C ratios and are Gram-positive (Fields 2004; Bik *et al*. 2006). Therefore, the phyla contain organisms that can transform PCE and TCE and contribute to their remediation at this site.

Research Question 1: Determine the effects of plant presence with regards to microbial diversity and dominance

Plant presence had an effect on microbial community composition and diversity. This outcome was expected based on the literature, but this research provided clear composition charts and richness and evenness parameters to support this hypothesis.

In order to address community composition, the sample sequences were compared to a known database, RDP, of 16S rRNA sequences and classified into phyla. Results from the RDP phyla classification showed that the organisms from the planted and control communities were classified into 17 and 19 phyla respectively. This included an Unclassified Bacteria category, which was reserved for any sample sequence that did not match a known sequence in the RDP database 80% or better. The control community had two phyla not seen in the planted community: TM7 and *Lentisphaerae*. TM7 is a candidate phylum that was named recently. The term candidate phylum refers to phyla-level clades with no cultured representatives, typically known only by limited numbers of rRNA sequences (Harris 2004). TM7 has been identified through its DNA, and has not yet been cultured, but a recent study shows that the phylum is widely distributed in the environment (Hugenholtz *et al*. 2001). TM7 was named after sequences obtained from a peat bog, activated sludge, and soil (Hugenholtz *et al*. 1998). The phylum *Lentisphaerae* is typically associated with marine organisms and has a strong relation to the phylum *Verrucomicrobia*. The phylum was discovered in 2004 in samples cultivated from Oregon coast seawater, and the species within the phylum are strictly aerobic (Cho 2004).

Since the sequences produced during this experiment were a small representative sample of the total microbial population, the presence of one individual in a phylum was important to document. The microbial community was extremely diverse here, and the individuals present in the sample represent the dominant organisms in the total community. All other phyla were present in approximately the same percentages for the planted and control communities; therefore, there were no other differences between the community composition of the planted and control communities to note. To verify that the communities were similar in composition, ANOSIM, a statistical similarity test, was performed on the RDP phylum classifications. The analysis revealed no significant differences between the planted and control communities, (n=5000 permutations; p=0.75). The microbial community composition did not change due to plant presence.

Community diversity was calculated using DOTUR, which compared sample sequences to one another and placed sequences into OTUs, based on sequence similarity. At the species level, 97% similarity, the evenness was high, while the sampling effort, according to Good's coverage, was low, an average of 45% for both the planted and control communities. This indicated that the species level was vastly undersampled. The true diversity was extremely high, which was the trend expected from literature on species microbial diversity. Species richness could not be determined because this level was undersampled. However, this data does support the accepted theory that the true microbial diversity in soil is extremely vast.

Good's coverage values indicated that the sampling effort for the phylum level was extremely high, ~90%, for both the planted and control communities. That, coupled with the fact that the evenness percentages at the phylum level were also high, illustrates that the phylum level diversity could be captured by the sample sequences. Richness parameters were significantly higher in the planted community compared to the control community. Communities associated with plant life are significantly more diverse than unplanted communities.

Research Question 2: Determine the effects of plant species with regards to microbial diversity and community composition

The results for this research showed that plant species produced different microbial composition in the mesocosms, but they were not significantly different. RDP phyla classifications illustrated some differences in the microbial communities associated with each species of plant. The *Firmicutes* population made up 3.4% of the total community in the *Carex comosa* mesocosms. While the *Firmicutes* population only reached 0.46% in the *Scirpus atrovirens* mesocosm and 1.2% in the *Eleocharis erythropoda*. Another difference was observed in the *Verrucomicrobia* population. *Eleocharis erythropoda* held the highest percentage with 4.4%, and the other two species had an average of 2.3%. This indicated that *Carex comosa* had a more prevalent population of *Firmicutes* in the microbial community associated with it. The last item to mention was that all three species of plants had a prevalent population of *Proteobacteria*. However, *Scirpus atrovirens* had nearly half of its individuals in this phylum while the other two communities only had a 33.7% makeup. This was expected since this phylum contains typical soil organisms. These differences illustrated that the plants can contribute to a microbial composition that was more prevalent to particular phyla. Previous studies have shown that different plant species can exude nutrients or other inputs that can affect the microbial community composition (Stottmeister 2003).

To verify that the community compositions were different, ANOSIM, a statistical similarity test, was performed on the RDP phylum classifications. The analysis revealed no significant differences between the plant species communities, (n=5000 permutations; p=0.21). Therefore even with the noted differences above, the community compositions were not significantly affected by the three plant species used in this experiment.

Diversity analysis was performed using DOTUR. The richness parameters showed some slight differences. At the species level, the evenness was high while the sampling effort, according to Good's coverage, was low, an average of 30% for all three communities. This indicated that the species level was vastly undersampled. The true diversity was extremely high, which was the trend expected from literature on microbial diversity.

At the phylum level the evenness was again high for all three communities and the sampling effort, according to Good's coverage, was also high, an average of 92%. At the phylum level, the rarefaction data for *Eleocharis erythropoda* was the highest for all three plant species. *Scirpus atrovirens* had species richness slightly below *Eleocharis erythropoda*, and *Carex comosa* had the lowest estimation. However, when 95% confidence intervals calculated by DOTUR were noted, this trend was not statistically significant. The *Eleocharis* and *Scirpus* communities overlapped error bars more than 50% as did the *Scirpus* and *Carex* communities. This indicated that the communities' richness were not statistically different. The *Carex* community did not overlap the *Eleocharis* community's error bars on the phylum level rarefaction curve, and therefore, the two communities' richness was statistically different. Therefore, the *Eleocharis erythropoda* had a more diverse community than *Carex comosa*. Also Figure 18, illustrated that with increased sampling effort the plant species phylum rarefaction curves will split apart and become significantly different for phylum richness.

Plants have been shown to increase diversity throughout the literature as well as above in Objective 1. However, plant species affect the microbial communities in various ways depending on the nutrients, root system, and other properties. The plants used in this research all came from the same family and have the same growth habit. Therefore, it was expected that the diversity and composition between the plant species would not differ. However, the results illustrate that the diversity for the *Carex* and *Eleocharis* communities do differ significantly. Therefore, there may be a metabolic property or other factor that one of the species had that affects the microbial community associated with it.

Research Question 3: Determine the effects of soil depth with regards to microbial diversity and community composition

There was evidence that microbial communities varied in composition due to depth. The depth communities represented the relationships established from a subsurface flow hydrology. RDP phyla classifications illustrated some differences in the microbial communities associated with depth. One phylum did stand out between the three depths. *Chloroflexi* was present at 4.4% in the bottom depth. The top and middle depths had only a 2.5% population. This could indicate that the bottom depths are more likely to promote an environment in which the phylum *Chloroflexi* can become prevalent *Chloroflexi* is a phylum that is known to contain dehalogenators. Dehalogenators are organisms that can bioremediate contaminants such as PCE and TCE, which are the contaminants treated by the WPAFB constructed wetland. To verify that the community compositions were different, ANOSIM, a statistical similarity test, was performed on the RDP phylum classifications. The analysis revealed no significant differences between the depth communities, (n=5000 permutations; p=0.31). Therefore even with the noted differences above, the community compositions are not significantly affected by depth in this study.

The diversity analysis was calculated using DOTUR. At the species level, the evenness was high while the sampling effort, according to Good's coverage, was low, an average of 35%. This indicated that the species level was vastly undersampled. The true diversity was extremely high, which was the trend expected from literature on microbial diversity.

Good's coverage values indicated that the sampling effort for the phylum level was extremely high, ~90%, for all the depth communities. The evenness at the phylum level was high indicating that the distribution of OTUs was even. Richness analysis did show that depth had an impact on richness at the phylum level. The *Carex* community and the control community richness were significantly lower in the middle layer than the other two depths. This indicates that these communities have a lower richness in the middle depth. However, all three species of plant communities and the control do show that with increased sampling depth richness does continue to vary and split apart from one another. The middle layer was consistently the lowest richness. This may be due to the fact that the middle layer was lacking or promoting nutrients, or other properties, that decrease diversity.

It is also interesting to note that the *Carex comosa* phyla rarefaction curve reached an asymptote for the middle layer lower than any of the other plant species or control communities, indicating that the *Carex comosa* community was associated with a lower diversity. As discussed in Chapter II, plant species can exude nutrients or have metabolic functions that are unique. These properties allow for a unique microbial community to form when associated with a particular plant species. Although *Carex comosa* is related to the other two plant species used in this study, the results presented here illustrate that it still has unique properties affecting the microbial community.

Limitations of research

This research was an attempt to characterize the soil microbial communities associated with plant presence, controls, and different plant species. Considering that a single gram of soil can potentially have 10^6 microorganisms, a sample size of 3,099 may be too small. However, reasonable interpretations can be made from the results of the sample. Another limitation involved the PCR amplification. In this project PCR amplified the 16S rRNA gene segment. This was in turn cloned. However, there is no guarantee that the clone generated from the PCR product was an original amplification or just another copy. Therefore, it should be mentioned that this analysis captures the dominant organisms within populations. Results should be interpreted within this context.

Also it is important to mention that the three species of plant chosen for this experiment share common ancestry and have the same herb growth habit. This means that the plants are not very different in how they operate, and therefore they would likely impact the microbial communities in a similar fashion. If diversity was the goal, it might have been more advantageous to use plants with different growth habits.

Significance of Research

This research was unique for several reasons. First, this analysis has never before been used with a mesocosm experiment. Studies using microcosms or field samples are common. Secondly 2,820 sequences were used for analysis. Previous research usually concentrated on ~100 to ~700 sequences. This research has increased the sample size four times. This allowed more complex results and interpretations. Lastly, this research is significant because it merged two detailed analyses together. The sequences were specifically classified into named phyla by the RDP program and then the sequences were grouped, based on evolutionary distances, using Phylip and DOTUR. This provided an in-depth analysis of the large amount of sequences generated by this project. The results provide invaluable insight into plant effect on microbial communities and depth effects. Most importantly, this research enhances the understanding of microbial consortia needed for bioremediation.

Further Research

This research simply hints at the true diversity of the microbial world. Therefore, it is recommended that further research is done to increase the sample size upwards to 8000 sequences. This sample size would be expected to approach the asymptote values seen in all the richness estimations in this research. Therefore the true diversity can be seen.

Also, since this research serves as a pre contamination baseline for comparison to PCE contaminated mesocosms, research should continue. This experiment should be repeated with samples from the now-contaminated mesocosms used for this experiment. This will allow researchers to determine the true effect PCE contamination has on microbial community composition and diversity. PCE contamination would be expected to affect the diversity and composition of the microbes in the mesocosms. Studies have shown that microbial communities change to handle specific contaminants. Therefore, it is hypothesized that the post contaminated samples will show less diversity and a stronger prevalence for phyla containing known dehalogenators and anaerobic organisms.

This research not only provides the baseline for comparison to contaminated sample, but it also provides a baseline to investigate the trends identified. This research showed that Chloroflexi had more prevalent in the bottom layers of all the mesocosm. This could indicate that the bottom layer had an environment more prone for organisms with this phylum. The first stages of remediation in a subsurface flow wetland occur in the bottom layers, and that was where the dehalogenators were expected. The *Carex comosa* community had a significantly lower richness at the middle level. This combined with other research illustrates that *Carex* has properties that diminish richness. An experiment should be organized to investigate this trend in *Carex*. And finally this baseline provided the composition makeup in the mesocosms. Now further research can investigate phyla and functional groups identified by this research using PCR specifically designed for identifying particular groups.

Summary

This research has shown some interesting trends in microbial communities that are most likely happening in the constructed wetland. The mesocosms were designed with the same soil properties, hydrologic flow, and plant presence. Therefore, the trends seen in the mesocosms are most likely also being experienced in the wetland at WPAFB. Microorganisms are an invaluable natural remediation system. Research such as this, provides the background understanding to help natural remediation become a more controlled and advantageous process.

Appendix A: PCR Protocol Using HotStarTaq Master Mix (Qiagen 2002).

This protocol serves only as a guideline for PCR amplification. Optimal reaction conditions, such as incubation times and temperatures, and amount of template DNA, may vary and need to be determined individually.

Notes:

- **Each PCR program should be started with an initial activation step of 15 min at 95°C to activate HotStarTaq DNA Polymerase (see step 6 of this protocol).**
- HotStarTaq Master Mix provides a final concentration of 1.5 mM $MgCl_2$ in the final reaction mix, which will produce satisfactory results in most cases. However, if a higher Mg_{2+} concentration is required, prepare a stock solution containing 25 mM $MgCl_2$.
- Set up reaction mixtures in an area separate from that used for DNA preparation or PCR product analysis.
- Use disposable tips containing hydrophobic filters to minimize cross-contamination.

1. Thaw primer solutions.

Mix well before use.

Optional: prepare a primer mix of an appropriate concentration (see Table 4) using the water provided. This is recommended if several amplification reactions using the same primer pair are to be performed. The final volume of diluted primer mix should be 25 µl per reaction including the template DNA, added at step 4.

2. Mix the HotStarTaq Masters Mix by vortexing briefly and dispense 25 µl into each PCR tube according to Table 4.

It is important to mix the HotStarTaq Master Mix before use in order to avoid localized concentrations of salt. HotStarTaq Master Mix is provided as a 2x concentrate (i.e., a 25µl volume of the HotStarTaq Master Mix is required for amplification reactions with a final volume of 50µl). For volumes smaller than 50 µl, the 1:1 ratio of HotStarTaq Master Mix to diluted primer mix and template should be maintained as defined in Table 4. A negative control (without template DNA) should always be included. It is not necessary to keep PCR tubes on ice as nonspecific DNA synthesis cannot occur at room temperature due to the inactive state of Hot StarTaq DNA Polymerase.

3. Distribute the appropriate volume of diluted primer mix into the PCR tubes containing the Master Mix.

4. Add template DNA (γ<=1 µg/reaction) to the individual PCR tubes.

The volume added should not exceed 10% of the final PCR volume.

Table 4. Reaction composition using HotStarTaq Master Mix

Component	Volume/reaction	Final concentration
HotStarTaq Master Mix 25 µl	25 µl	2.5 units HotStarTaq DNA Polymerase 1 x PCR Buffer* 200 µM of each dNTP
Diluted primer mix		
Primer A 0.1-0.5 µM	Variable	0.1-.05 µM
Primer B	Variable	0.1-.05 µM
Distilled water (provided)	Variable	-
Template DNA		
Template DNA, added at step 4	Variable	≤1 µg/reaction
Total Volume	50 µl	-

*Contains 1.5 mM $MgCl_2$

5. When using thermal cyclers with a heated lid, do not use material oil. Proceed directly to step 6. Otherwise, overlay with approximately 50 µl mineral oil.

6. **Program the thermal cycler according to the manufacturer's instructions.** Each PCR program must start with an initial heat activation step at 95ºC for 15 min. A typical PCR cycling program is outlined below. For maximum yield and specificity, temperatures and cycling times should be optimized for each new template target and primer pair.

			Additional Comments
Initial activitation step:	**15 min**	**95°C**	**HotStarTaq DNA Polymerase is** activated by this heating step
3-step cycling			
Denaturation:	**0.5-1 min**	**94°C**	
Annealing:	**0.5-1 min**	**50°C-68°C**	**5°C below Tm of primers**
Extension:	**1 min**	**72°C**	**For PCR products longer than 1kb,** use an extension time of approximately 1 min per kb DNA
Number of Cycles:	**20-35**		
Final Extension:	**10 min**	**72°C**	

7. Place the PCR tubes in the thermal cycler and start cycling program. Note: After amplification, samples can be stored overnight at 2-8ºC or at -20ºC for longer storage.

Appendix B: Cloning Month Legend

A-August
S-September
O-October
N-November
D-December
J-January
F-February
M-March
Ap-April
My-May
Ju-June
Jy-July

Appendix C: Mo Bio PowerSoil™ DNA Isolation Kit Extraction Protocol. (Mo Bio Laboratories, Carlsbad, CA 2004)

Introduction

The PowerSoil™ DNA Isolation Kit is comprised of a novel and proprietary method for isolating genomic DNA from environmental samples. The kit is intended for use with environmental samples containing a high humic acid content including difficult soil types such as compost, sediment, and manure. Other more common soil types have also been used successfully with this kit. The isolated DNA has a high level of purity allowing for more successful PCR amplification of organisms from the sample. PCR analysis has been performed to detect a variety of organisms including bacteria (*e.g. Bacillus subtilis, Bacillus anthracis*), fungi (*e.g.* yeasts , molds), algae and Actinomycetes (*e.g. Streptomyces*).

The PowerSoil™ DNA Isolation Kit distinguishes itself from Mo Bio's Ultraclean™ Soil DNA Isolation kit with a **NEW** humic substance/brown color removal procedure. This new procedure is effective at removing PCR inhibitors from even the most difficult soil types.

Environmental samples are added to a bead beating tube for rapid and thorough homogenization. Cell lysis occurs by mechanical and chemical methods. Total genomic DNA is captured on a silica membrane in a spin column format. DNA is then washed and eluted from the membrane. DNA is then ready for PCR analysis and other downstream applications.

WARNING: Solution C5 contains ethanol. It is flammable.

IMPORTANT NOTE FOR USE: Make sure the 2 ml PowerBead Tubes rotate freely in your centrifuge without rubbing.

Kit Storage

Kit reagents and components should be stored at room temperature.

Kit Contents

	Quantity	
Component	**12888-50**	**12888-100**
PowerBead Tubes (contains 750 ul solution	50	100
Solution C1	3.3 ml	6.6 ml
Solution C2	14 ml	28 ml
Solution C3	11 ml	22 ml
Solution C4	72 ml	144 ml
Solution C5	27.5 ml	55 ml
Solution C6	6 ml	12 ml
Spin Filters Units in 2 ml Tubes	50	100
Collection Tubes (2 ml)	200	400

1. To the 2 ml PowerBead Tubes provided, add 0.25 gm of soil sample.
2. Gently vortex to mix.
3. **Check solution C1.** If Solution C1 is precipitated, heat solution to 60ºC until dissolved before use.
4. Add 60 µl of Solution C1 and invert several times or vortex briefly.
5. Secure PowerBead Tubes horizontally using the Mo Bio Vortex Adapter tube holder for the vortex (Mo Bio Catalog No. 13000-V1. Call 1-800-606-6246 for information) or secure tubes horizontally on a flat-bed vortex pad with tape. Vortex at maximum speed for 10 minutes.
6. Make sure the PowerBead Tubes rotate freely in your centrifuge without rubbing. Centrifuge tubes at 10,000 x *g* for 30 seconds. **CAUTION:** Be sure not to exceed 10,000 x *g* or tubes may break.
7. Transfer the supernatant to a clean microcentrifuge tube (provided).
   Note: Expect between 400 to 500 µl of supernatant. Supernatant may still contain some soil particles.
8. Add 250 µl of Solution C2 and vortex for 5 seconds. Incubate at 4ºC for 5 minutes.
9. Centrifuge the tubes for 1 minute at 10,000 x *g*.
10. Avoiding the pellet, transfer up to, but no more than, 600µl of supernatant to a clean microcentrifuge tube (provided).
11. Add 200µl of Solution C3 and vortex briefly. Incubate at 4ºC for 5 minutes.
12. Centrifuge the tubes for 1 minute at 10,000 x *g*.
13. Avoiding the pellet, transfer up to, but no more than, 750µl of supernatant to a clean microcentrifuge tube (provided).
14. Add 1200 µl of Solution C4 to the supernatant and vortex for 5 seconds.
15. Load approximately 675 µl onto a spin filter and centrifuge at 10,000 x *g* for 1 minute. Discard the flow through and add an additional 675 µl of supernatant to the spin filter and centrifuge at 10,000 x *g* for 1 minute. Load the remaining supernatant onto the

spin filter and centrifuge at 10,000 x *g* for 1 minute. **Note:** A total of three loads for each sample processed are required.
16. Add 500 µl of Solution C5 and centrifuge for 30 seconds at 10,000 x *g*.
17. Discard flow through.
18. Centrifuge again for 1 minute.
19. Carefully place spin filter in a new clean tube (provided). Avoid splashing any Solution C5 onto the spin filter.
20. Add 100µl of Solution C6 to the center of the white filter membrane. Alternatively, sterile DNA-Free PCR Grade Water may be used for elution from the silica spin filter membrane at this step (Mo Bio Catalog No. 17000-10).
21. Centrifuge for 30 seconds.
22. Discard the spin filter. DNA in the tube is now application ready. No further steps are required.

We recommend storing DNA frozen (-20ºC to -80ºC). Solution C6 contains no EDTA.

Wet Soil Sample

If soil sample is high in water content, remove contents from PowerBead Tube (beads and solution) and transfer into another sterile microcentrifuge tube (not provided). Add soil sample to PowerBead Tube and centrifuge for 30 seconds at 10,000 x *g*. Remove as much liquid as possible with a pipet tip. Add beads and bead solution back to PowerBead Tube and follow protocol starting at step 2.

If DNA Does Not Amplify

- Make sure to check DNA yields by gel electrophoresis or spectrophotometer reading. An excess amount of DNA will inhibit PCR reaction.
- Diluting the template DNA should not be necessary with DNA isolated with the PowerSoil DNA Isolation Kit; however, it should still be attempted.
- If DNA will still not amplify after trying the steps above, then PCR optimization (changing reaction conditions and primer choice) may be needed.

Eluted DNA Sample Is Brown

We have not observed any coloration in DNAs isolated using the PowerSoil DNA Isolation Kit. If you observe coloration in your samples, please contact technical support for suggestions.

Alternative Lysis Method

After adding Solution C1, vortex 3-4 seconds, then heat to 70ºC for 5 minutes. Vortex 3-4 seconds. Heat another 5 minutes. Vortex 3-4 seconds. This alternative procedure will reduce shearing but may also reduce yield.

Concentrating the DNA

Your final volume will be 100µl. If this is too dilute for your purposes, add 4 µl of 5M NaCl and mix. Add 200 µl of 100% cold ethanol and mix. Centrifuge at 10,000 x *g* for 5 minutes. Decant all liquid. Dry residual ethanol in a speed vac, dessicator, or air dry. Resuspend precipitated DNA in desired volume.

DNA Floats Out of Well When Loaded on a Gel

You may have inadvertently transferred some residual Solution C5 into the final sample. Prevent this by being careful in step19 not to transfer liquid onto the bottom of the spin filter basket. Ethanol precipitation is the best way to remove Solution C5 residue. (See "Concentrating the DNA" above)

Storing DNA

DNA is eluted in Solution C6 (10mM Tris) and must be stored at -20ºC to 80ºC or it may degrade over time. DNA can be eluted in TE but the EDTA may inhibit reactions such as PCR and automated sequencing. DNA may be eluted with sterile DNA-Free PCR Grade Water (Mo Bio Catalog No. 17000-10).

Cells are Difficult to Lyse

If cells are difficult to lyse, a 10 minute incubation at 70ºC, after adding Solution C1, can be performed. Follow by continuing with protocol step 5.

Technical Information

Product Manuafactured by Mo Bio Laboratories, Inc. 2746 Loker Avenue West, Carlsbad, CA 92008.

Appendix D: StrataClone™ PCR Cloning Kit

MATERIALS PROVIDED

	Quantity[a]	
Materials Provided	Catalog # 240205	Catalog # 240206
StrataClone™Vector Mix	21 reactions (µl each)	
StrataClone™Cloning Buffer	63 µl	63 µl
StrataClone™Control Insert (5 ng/µl)	50 ng	50 ng
StrataClone™SoloPack®Competent Cells	21 transformations (50 µl each)	11 transformations (50 µl each)
pUC18 Control Plasmid (0.1 ng/µl in TE Buffer)	10 µl	10 µl

[a] Catalog #240205 provides enough reagents for 20 experimental cloning reactions plus one Control Insert cloning reaction. Catalog #240206 provides enough reagents for 10 experimental cloning reactions plus one Control Insert cloning reaction.

STORAGE CONDITIONS

StrataClone™ SoloPack® Competent Cells and pUC18 Control Plasmid: –80°C
All Other Components: –20°C

Note The StrataClone SoloPack competent cells are sensitive to variations in temperature and must be stored at the bottom of a –80°C freezer. Transferring tubes from one freezer to another may result in a loss of efficiency.

ADDITIONAL MATERIALS REQUIRED

Taq DNA polymerase or a polymerase blend recommended for PCR cloning
Thermocycler
LB–ampicillin agar plates
LB medium
5-Bromo-4-chloro-3-indoyl-β-D-galactopyranoside (X-gal)

INTRODUCTION

The StrataClone™ PCR Cloning Kit§ allows high-efficiency, 5-minute cloning of PCR products, using the efficient DNA rejoining activity of DNA topoisomerase I and the DNA recombination activity of Cre recombinase.

Overview of StrataClone™ PCR Cloning Technology

StrataClone PCR cloning technology exploits the combined activities of topoisomerase I from *Vaccinia* virus and Cre recombinase from bacteriophage P1. *In vivo*, DNA topoisomerase I assists in DNA replication by relaxing and rejoining DNA strands. Topoisomerase I cleaves the phosphodiester backbone of a DNA strand after the sequence 5′-CCCTT, forming a covalent DNA–enzyme intermediate which conserves bond energy to be used for religating the cleaved DNA back to the original strand. Once the covalent DNA–enzyme intermediate is formed, the religation reaction can also occur with a heterologous DNA acceptor.1 The Cre recombinase enzyme catalyzes recombination between two loxP recognition sequences.

The StrataClone PCR cloning vector mix contains two DNA arms, each charged with topoisomerase I on one end and containing a loxP recognition sequence on the other end. The topoisomerase-charged ends have a modified uridine (U*) overhang. Taq-amplified PCR products, which contain 3′-adenosine overhangs, are efficiently ligated to these vector arms in a 5-minute ligation reaction, through A-U* base-pairing followed by topoisomerase I-mediated strand ligation.

The resulting linear molecule (vector arm^{ori}–PCR product–vector arm$^{amp)}$ is then transformed, with no clean-up steps required, into a competent cell line engineered to transiently express Cre recombinase. Cre-mediated recombination between the vector loxP sites creates a circular DNA molecule (pSC-A-amp/kan, see Figure 2) that is proficient for replication in cells growing on media containing ampicillin. The resulting pSC-A product includes a *lac*Z′ α-complementation cassette for blue-white screening.

StrataClone™ SoloPack® Competent Cells

The provided StrataClone SoloPack competent cells express Cre recombinase, in order to circularize the linear DNA molecules produced by topoisomerase I-mediated ligation. The cells are provided in a convenient single-tube transformation format. This host strain (containing the lacZΔM15 mutation) supports blue-white screening with plasmid pSC-A, containing the *lac*Z′ α-complementation cassette (see Figure 2). It is not necessary to induce *lac*Z′ expression with IPTG when performing blue-white screening with this strain.

The StrataClone SoloPack competent cells are optimized for high efficiency transformation and recovery of high-quality recombinant DNA. The cells are endonuclease (*end*A), and recombination (*rec*A) deficient, and are restriction-minus. The cells lack the tonA receptor, conferring resistance to T1, T5, and φ80 bacteriophage infection, and lack the F′ episome. StrataClone SoloPack competent cells are resistant to streptomycin.

PCR CLONING PROTOCOL

Preparing the PCR Product

1. Prepare insert DNA by PCR using Taq DNA polymerase or an enzyme blend qualified for PCR cloning applications.

Note Taq DNA polymerase is required for the addition of 3′-adenine residues to the PCR product. If PCR was performed using a proofreading DNA polymerase, see Appendix II for a protocol for adding 3′-A overhangs after the PCR reaction is complete.

If the PCR template is a plasmid encoding the ampicillin resistance gene, the plasmid DNA must be eliminated prior to the cloning reaction by Dpn I digestion or by gel purification of the PCR product.

2. Analyze an aliquot of the PCR reaction on an agarose gel to verify production of the expected fragment.

3. If the fragment to be cloned is <3 kb and gel analysis confirms robust, specific amplification, prepare a 1:10 dilution of the PCR reaction in dH_20. For larger or poorly amplified fragments, omit the dilution step.

Note If multiple PCR products are observed on the gel, or when cloning very large PCR products, gel isolate the desired PCR product prior to performing the ligation reaction. See Appendix I for a gel-isolation protocol. For a gel-isolated PCR product recovered in 50 µl, add 2 µl (undiluted) of the purified PCR product to the ligation reaction below.

Ligating the Insert

4. Prepare the ligation reaction mixture by combining (in order) the
following components:
3 µl StrataClone™ Cloning Buffer
2 µl of PCR product (5–50 ng, typically a 1:10 dilution of a robust PCR reaction)
or 2 µl of StrataClone™ Control Insert
2 µl StrataClone™ Vector Mix

5. Mix gently by repeated pipetting, and then incubate the ligation reaction at room temperature for 5 minutes. When the incubation is complete, place the reaction on ice.

Note The cloning reaction may be stored at –20°C for later processing.

Transforming the Competent Cells

6. Thaw one tube of StrataClone SoloPack competent cells on ice for each ligation reaction.

Note It is critical to use the provided StrataClone SoloPack competent cells, expressing Cre recombinase, for this protocol. Do not substitute with another strain.

7. Add 1 µl of the cloning reaction mixture to the tube of thawed competent cells. Mix gently (do not mix by repeated pipetting).

> **Notes** For large PCR products, up to 2 µl of the cloning reaction mixture may be added to the transformation reaction.
>
> If desired, test transformation efficiency of the competent cells by transforming a separate tube of competent cells with 10 pg of pUC18 control DNA. Prior to use, dilute the pUC18 DNA provided 1:10 in dH20, and then add 1 µl of the dilution to the tube of competent cells.

8. Incubate the transformation mixture on ice for 20 minutes. During the incubation period, pre-warm SOC medium to 42°C.

9. Heat-shock the transformation mixture at 42°C for 45 seconds.

10. Incubate the transformation mixture on ice for 2 minutes.

11. Add 250 µl of pre-warmed SOC medium to the transformation reaction mixture. Allow the competent cells to recover for at least 1 hour at 37°C with agitation. (Lay the tube of cells on the shaker horizontally for better aeration.)

12. During the outgrowth period, prepare LB–ampicillin plates for blue-white color screening by spreading 40 µl of 2% X-gal on each plate.

13. Plate 5 µl and 100 µl of the transformation mixture on the LB–ampicillin-X-gal plates. Incubate the plates overnight at 37°C.

> Notes For the Control Insert cloning reaction, plate 10 µl of the transformation mixture.
>
> For the pUC18 control transformation, plate 30 µl of the transformation mixture.
>
> When spreading <50 µl of transformation mixture, pipette the cells into a 50-µl pool of SOC medium before spreading.

14. Pick white for plasmid DNA analysis.

> Notes Colonies harboring plasmids containing typical PCR product inserts are expected to be white. After prolonged incubation, some of the insert-containing colonies may appear light blue.

Appendix E: Plasmid Prep Protocol

QIAprep 8 Turbo Miniprep Kit Catalog no.	**(10) 27152**	**(50) 27154**
Turbofilter® 8 Strips	10	50
QIAprep 8 Strips	10	50
Buffer P1	40 ml	125 ml
Buffer P2	40 ml	125 ml
Buffer N3*	60 ml	2 x 125 ml
Buffer PB *	100 ml	500 ml
Buffer PE (concentrate)	2 x 20 ml	2 x 100 ml
Buffer EB	55 ml	2 x 55 ml
Rnase A	400 µl [t]	125 µl [T]
Collection Microtubes (1.2 ml)	13 x 8	55 x 8
Caps for QIAprep Strips	13 x 8	55 x 8
Caps for Collection Microtubes	13 x 8	55 x 8
Handbook	1	1

* Buffers N3 and PB contain Chaotrophic salts which are irritants and not compatible with disinfecting agents containing bleach. Take appropriate laboratory safety measures and wear gloves when handling.

[t] Provided as a 10 mg/ml solution

[T] Provided as a 100 mg/ml solution

Introduction

The QIAprep Miniprep system provides a fast, simple, and cost-effective plasmid miniprep method for routine molecular biology laboratory applications. QIAprep Miniprep Kits use silica membrane technology to eliminate the cumbersome steps associated with loose resisns or slurries. Plasmid DNA purified with QIAprep Miniprep Kits is immediately ready for use. Phenol extraction and ethanol precipitation are not required, and high-quality plasmid DNA is eluted in a small volume of Tris buffer (included in each kit) or water. The QIAprep system consists of four products with different handling options to suit every throughput need.

Low throughput

The **QIAprep Spin Miniprep Kit** is designed for quick and convenient processing of 1-24 samples simultaneously in less than 30 minutes. QIAprep spin columns can be used in a microcentrifuge or on any vacuum manifold with luer connectors (e.g., QIAvac 24 Plus, or QIAvac 6S with QIAvac Luer Adapters).

Principle

The QIAprep miniprep procedure is based on alkaline lysis of bacterial cells followed by adsorption of DNA onto silica in the presence of high salt. The unique silica membrane used in QIAprep Miniprep Kit completely replaces glass or silica slurries for plasmid minipreps.

The procedure consists of three basic steps:

- Preparation and clearing of a bacterial lysate
- Adsorption of DNA onto the QIAprep membrane
- Washing and elution of plasmid DNA

Protocol: Plasmid DNA Purification Using the QIAprep Spin Miniprep Kit and a Microcentrifuge

This protocol is designed for purification of up to 20 µg of high-copy plasmid DNA from 1-5 ml overnight cultures of *E. coli* in LB (Luria-Bertani) medium.

1. **Resuspend pelleted bacterial cells in 250 µl Buffer P1 and transfer to a microcentrifuge tube.**
   Ensure that RNase A has been added to Buffer P1. No cell clumps should be visible after resuspension of the pellet.
   If LyseBlue reagent has been added to Buffer P1, vigorously shake the buffer bottle to ensure LyseBlue particles are completely dissolved. The bacteria should be resuspended completely by vortexing or pipetting up and down until no cell clumps remain.
2. **Add 250 µl Buffer P2 and mix thoroughly by inverting the tube 4-6 times.**
   Mix gently by inverting the tube. Do not vortex, as this will result in shearing of genomic DNA. If necessary, continue inverting the tube until the solution becomes viscous and slightly clear. Do not allow the lysis reaction to proceed for more than 5 min.
   If LyseBlue has been added to Buffer P1 the cell suspension will turn blue after addition of Buffer P2. Mixing should result in a homogeneously colored suspension. If the suspension contains localized colorless regions or if brownish cell clumps are still visible, continue mixing the solution until a homogeneously colored suspension is achieved.
3. **Add 350 µl Buffer N3 and mix immediately and thoroughly by inverting the tube 4-6 times.**
   To avoid localized precipitation, mix the solution thoroughly, immediately after addition of Buffer N3. Large culture volumes (e.g., ≥5 ml) may require inverting up to 10 times. The solution should become cloudy.
   If LyseBlue reagent has been used, the suspension should be mixed until all trace of blue has gone and the suspension is colorless. A homogeneous colorless suspension indication that the SDS has been effectively precipitated.
4. **Centrifuge for 10 min at 13,000 rpm (~17,900 x g) in a table-top microcentrifuge.**
   A compact white pellet will form.
5. **Apply supernatants from step 4 to the QIAprep spin columns by decanting or pipetting.**
6. **Centrifuge for 30-60 s. Discard the flow-through.**
7. **Wash the QIAprep spin column by adding 0.5 ml Buffer PB and centrifuging for**

30-60 s. Discard the flow-through.

8. **Wash QIAprep spin column by adding 0.75 ml Buffer PE and centrifuging for 30-60 s.**
9. **Discard the flow-through, and centrifuge for an additional 1 min to remove residual wash buffer.**

   Important: Residual wash buffer will not be completely removed unless the flow-through is discarded before this additional centrifugation. Residual ethanol from Buffer PE may inhibit subsequent enzymatic reactions.
10. **Place the QIAprep column in a clean 1.5 ml microcentrifuge tube. To elute DNA, add 50 µl Buffer EB (10mM Tris-Cl, pH 8.5) or water to the center of each QIAprep spin column, let stand for 1 min, and centrifuge for 1 min.**

Appendix F: Restriction Digest Protocol (Promega, Madison, WI 2008)

Introduction

Restriction enzymes, also referred to as restriction endonucleases, are enzymes which recognize short, specific (often palindromic) DNA sequences. They cleave double-stranded DNA (dsDNA) at specific sites within or adjacent to their recognition sequences. Most restriction enzymes (REs) will not cut DNA that is **methylated** on one or both strands of their **recognition site**, although some require substrate methylation.

Each restriction enzyme has specific requirements to achieve optimal activity. Ideal storage and assay conditions favor the most activity and highest fidelity in a particular enzyme's function. Conditions such as temperatures, pH, enzyme cofactor(s), salt composition and ionic strength affect enzyme activity and stability. Two buffers usually accompany each of the Promega's restriction enzymes. One buffer is the optimal reaction buffer which may be from the **4-CORE® System** (Reaction Buffers A, B, C, D) or one of the other optimal buffers (Reaction Buffers E-L), and the other is the MULTI-CORE™ Buffer. The supplied optimal buffer always yields 100% activity for the enzyme it accompanies, and serves as the specific reaction buffer for individual digests with that enzyme. **The MULTI-CORE™ Buffer**, which is designed for broad compatibility with many REs, is provided with enzymes that have 25% or greater activity in the buffer. The MULTI-CORE™ Buffer is useful for multiple digests because it generally yields more activity for more enzyme combinations than any of the other buffers, but sometimes with a compromise in activity. Multiple digests using REs with significantly different buffer requirements may require a sequential reaction with the addition of RE buffer or salt before the second enzyme is used

DNA Substrate Considerations

DNA substrates commonly used for restriction enzyme digestion include DNA from bacteriophage lambda, bacterial plasmid DNA and genomic DNA. Lambda DNA is a linear DNA form that is an industry standard for measuring and expressing unit activity for many restriction enzymes. Compared to linear DNA, intact supercoiled plasmid DNA (and DNAs with a large number of the target restriction site) required more units of enzyme (two- to tenfold) per microgram than the DNA used in the enzyme's activity assay.

PCR products and oligonucleotides are relatively small compared with DNA used for defining RE units. Therefore, when using these substrates in a restriction digest, it is essential to take into consideration the molar concentration of enzyme recognition sites and not just the mass DNA. Also, some REs require flanking bases surrounding the core RE restriction site. This is problematic when it is necessary to cut an oligonucleotide or a fragment of DNA with an RE site near its end. When PCR cloning strategies include the use of primers containing an RE site, care is necessary in designing the primer with adequate DNA surrounding the core RE recognition sequence.

In addition to the form and original source of the DNA, the purity is another factor that must be considered. Depending on the purification method and the handling of the DNA, it may contain varying amounts of contaminants that affect restriction enzyme digestion and analysis. Contaminants may include other types of DNA, nucleases, salts and inhibitors or restriction enzymes. The effect of a contaminant on an RE digest is generally dose-dependent: i.e., the inhibitory effects will increase with the volume of DNA added to the restriction enzyme reaction. Relatively pure DNA is required for efficient restriction enzyme digestion. Contaminating nucleases are usually activated only after the addition of salts (e.g., restriction enzyme buffer) to the DNA solution. Therefore, appropriate control reactions should always be run in parallel with the restriction digest. Buffer solutions containing EDTA in low concentrations (1mM) are often used to protect DNA from nuclease degradation during storage, but the EDTA can interfere with restriction enzyme digestion if the final concentration of EDTA in the reaction is too high. This situation usually results when the concentration of the substrate DNA is low and it is necessary to use a large volume of DNA in the digest. In such cases, it is best to concentrate the DNA (e.g., by ethanol precipitation). The organic solvents, salts, detergents and chelating agents that are sometimes used during the purification of DNA can also interfere with restriction enzyme activity if they carry over the final DNA solution. Dialysis and/or ethanol precipitation with 2.5 M ammonium acetate (final concentration before adding ethanol) followed by drying and resuspension can remove many of these substances. While relatively pure DNA is required for efficient for efficient restriction enzyme digestion, additional of acetylated BSA to a final concentration of 0.1 mg/ml can sometimes improve the quality and efficiency of enzyme assays containing impure DNA and we recommend that it be included in all digests.

Enzyme Storage, Handling and Use

Maintain the sterility of reagents used in the RE digest as well as any tools (e.g., tubes, pipette tips) used with those reagents. Restriction enzymes should be stored in a non-frost-free freezer, except for a brief period during use, when they should be kept on ice. The restriction enzyme is usually the last reagent added to a reaction, to ensure that it is not exposed to extreme conditions. When many similar digests are being prepared, it may be convenient to create premixes of common reagents.

Before assembling the restriction digest, thoroughly mix each component to be added to the reaction and then centrifuge the tubes of reagents briefly to collect the contents in the bottom of the tube. The reaction components should also be mixed after addition of the enzyme to the digest. While high salt buffers and glycerol-containing reagents are difficult to mix, all solutions containing restriction enzymes must be mixes gently to avoid inactivating the enzyme.

Setting up a Restriction Enzyme Digest (adapted from Promega protocol)

An analytical scale restriction enzyme digest is usually performed in a volume of 20 µl on 0.2-1.5 µg of substrate DNA, using a two- to tenfold excess of enzyme over DNA. If an unusually large volume of DNA or enzyme is used, aberrant results may occur and may not be readily recognized. The following is the protocol followed for this research:

1. Turn 37ºC water bath

2. Put BSA, Buffer H, *Eco*R1 on ice to thaw. Put DNA from selected samples in a tube holder to thaw.

3. Add ingredients one at a time as follows to an eppendorf tube. Don’t forget to label tube by sample and denote it is a restriction digest by adding RD to the label.

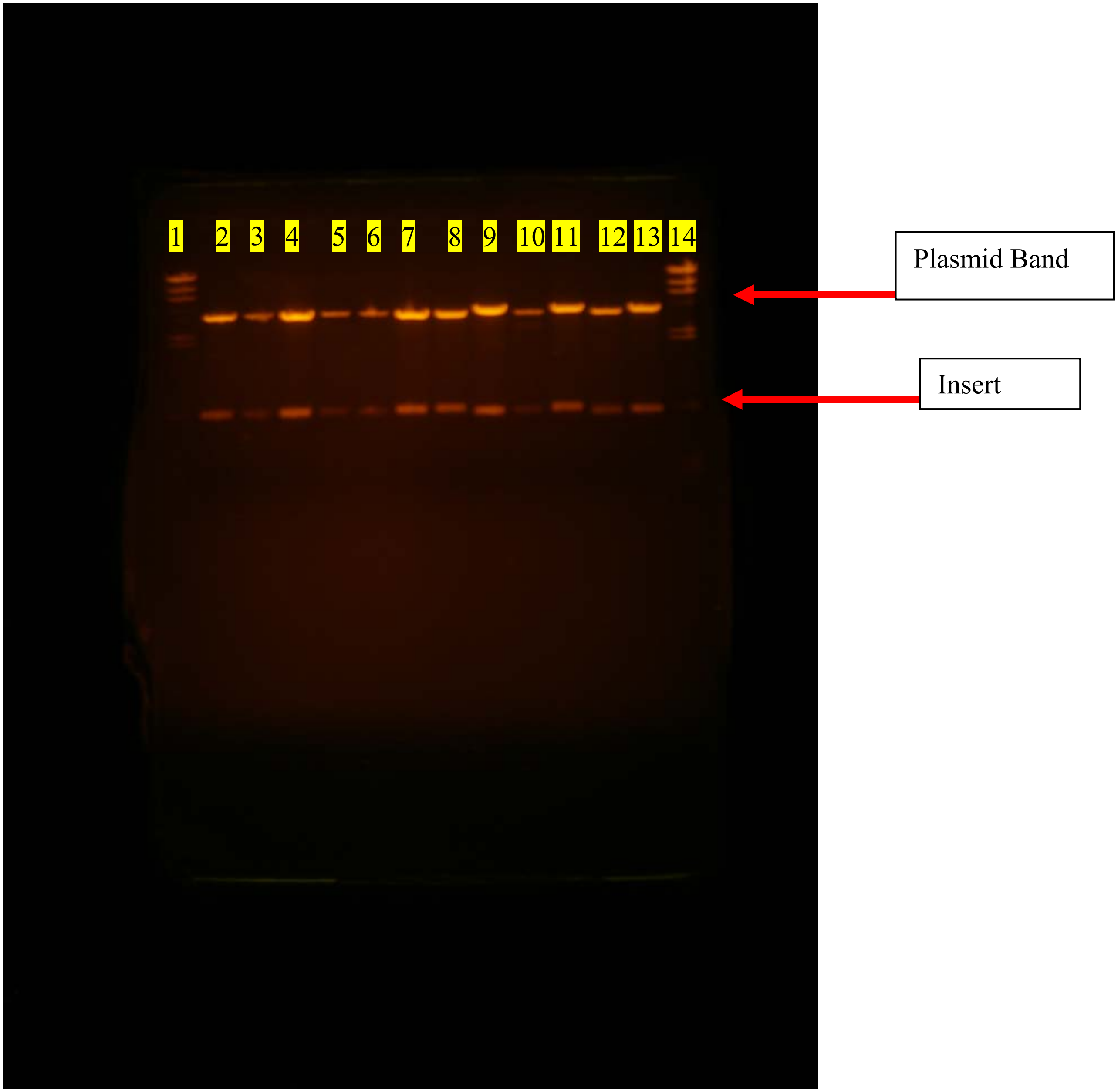
> 14.3 µl distilled water
> 2.0 µl Buffer H
> 3.0 µl DNA
> .2 µl BSA
> .5 µl *Eco*R1
> Total Volume 20 µl

4. Place all restricted digested samples in the water bath for 2-3 hours.

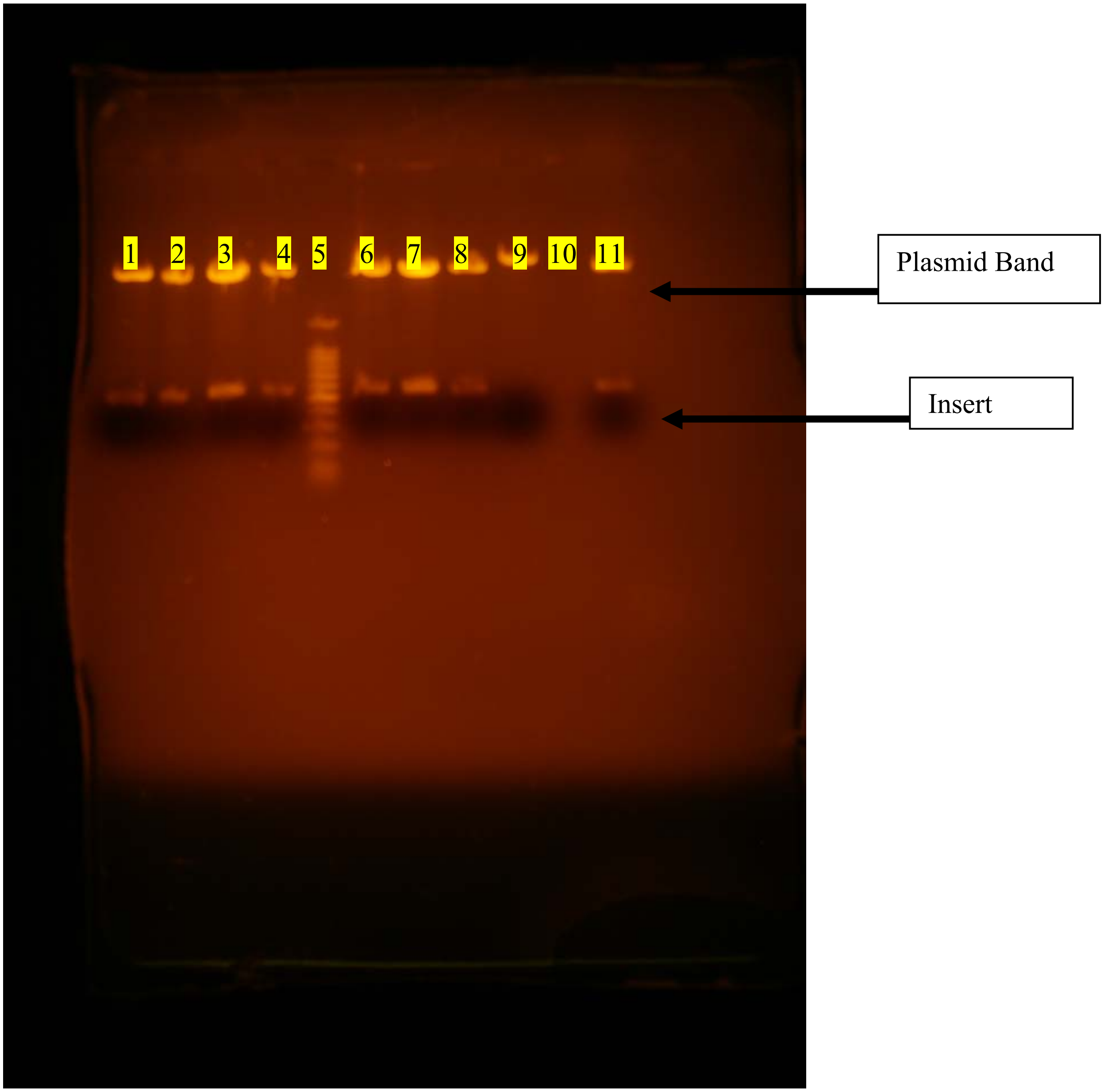
Experimental Controls

Experimental controls are necessary to identify, understand and explain problems or inconsistencies in results. The following controls are commonly used in parallel with RE digests: (i) uncut experimental DNA, (ii) digest of commercially supplied control DNA, (iii) no-enzyme “mock” digest, (iv) 1 of 2 different sizes markers in more than one lane per gel (i.e., different locations).

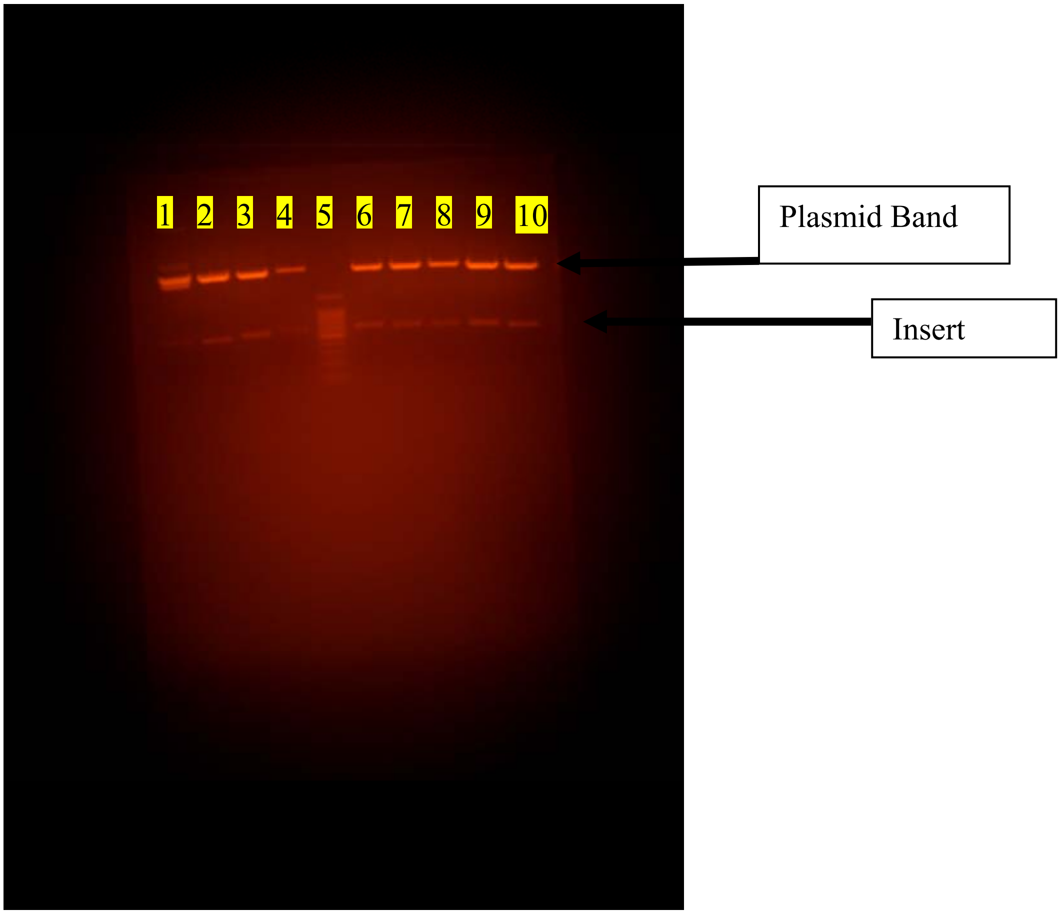
Appendix G: Gels



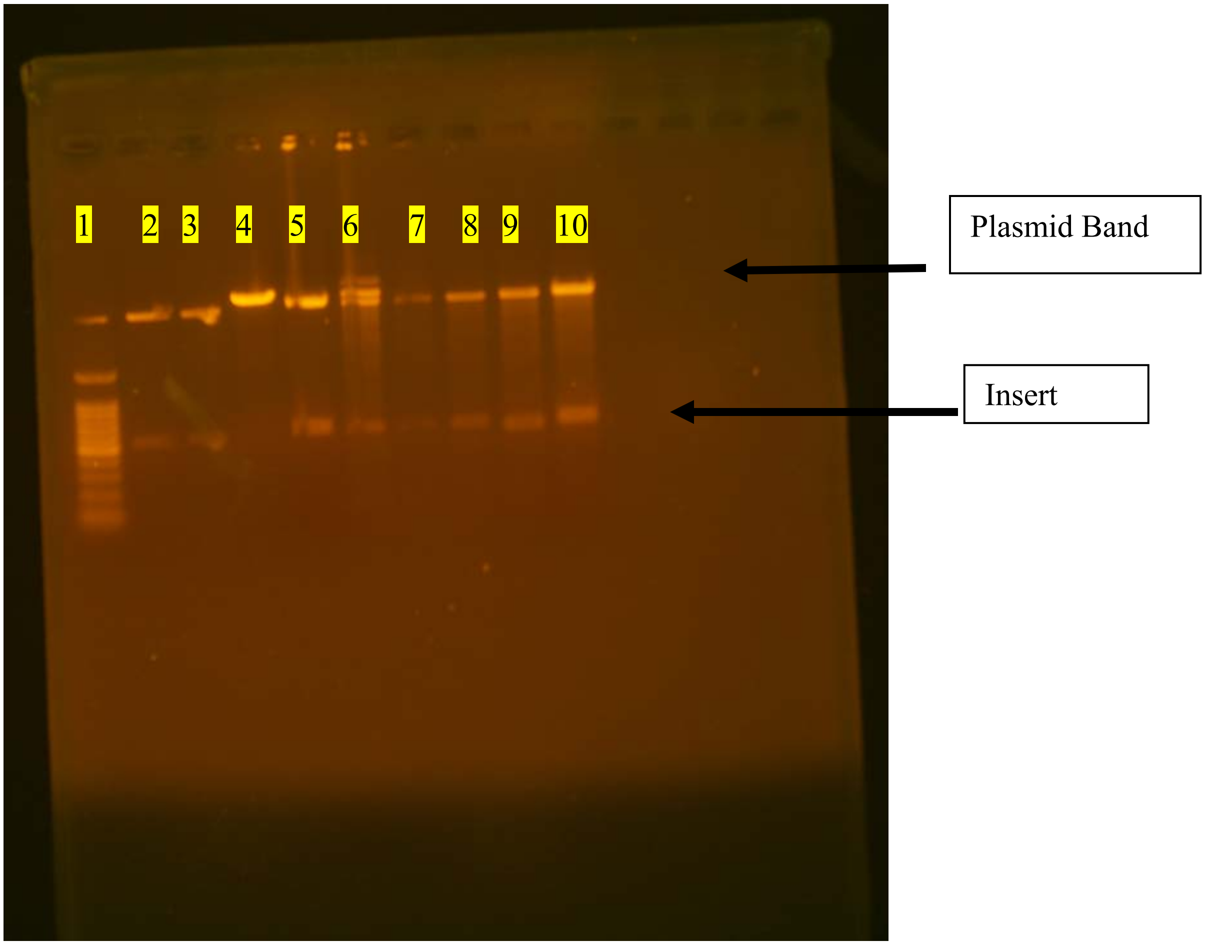
F11.L1 Gel: Lane 1-100bp ladder; Lane 2-F11.L1.5.24; Lane 3-F11.L1.6.12; Lane 4-F11.L1.1.24; Lane 5-F11.L1.1.36; Lane 6-F11.L1.3.23; Lane 7-F11.L1.1.21; Lane 8-F11.L1.3.24; Lane 9-A21.3.10; Lane 10-F11.L1.2.26; Lane 11-A21.3.21; Lane 12-F11.L1.2.22; Lane 13-A21.3.23; Lane 14-100 bp ladder



M11-1.L1 Gel: Lane 1-M11-1.L1.1.1; Lane 2-M11-1.L1.1.4; Lane 3-M11-1.L1.1.11; Lane 4-M11-1.L1.2.3; Lane 5-100 bp ladder; Lane 6-M11-1.L1.2.16; Lane 7-M11-1.L1.3.8; Lane 8-M11-1.L1.4.2; Lane 9-M11-1.L1.4.4; Lane 10-Empty; Lane 11-M11-1.L1.4.16

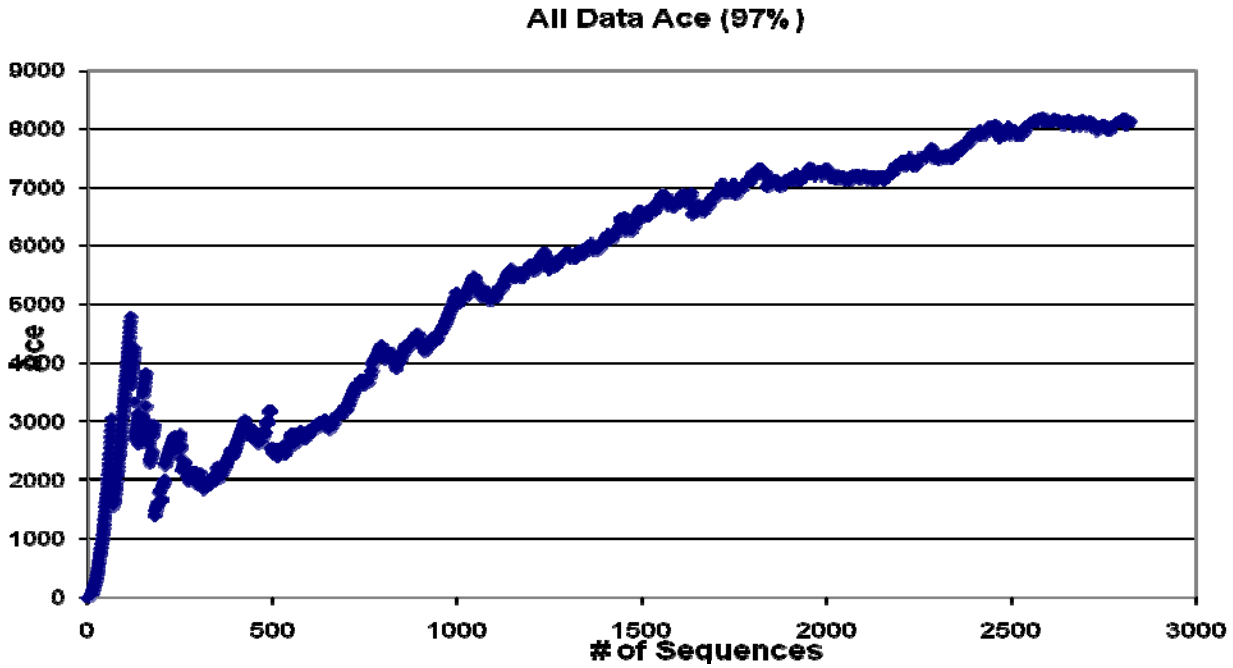


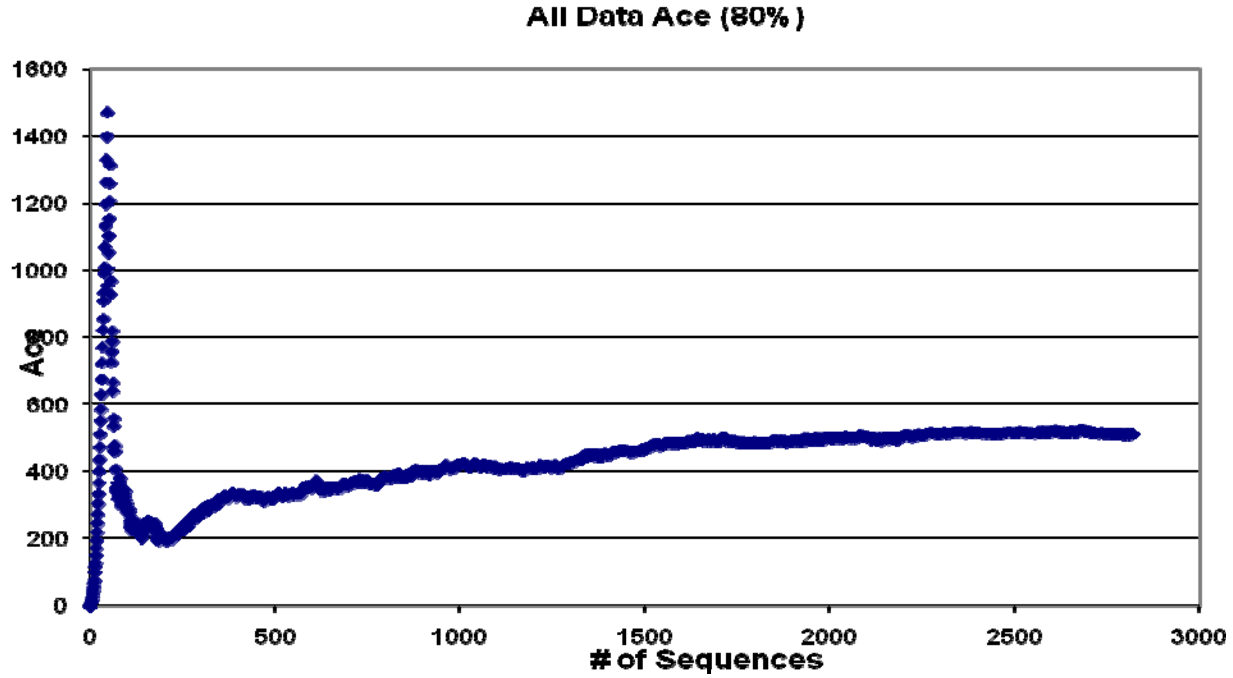
Ap53.L1 Gel: Lane 1-Ap53.L1.5.6; Lane 2-Ap53.L1.3.18; Lane 3-Ap53.L1.3.14; Lane 4-Ap53.L1.3.10; Lane 5: 100bp ladder; Lane 6-Ap53.L1.2.5; Lane 7-Ap53.L1.3.2; Lane 8-Ap53.L1.5.18; Lane 9-Ap53.L1.5.14; Lane 10-Ap53.L1.5.13

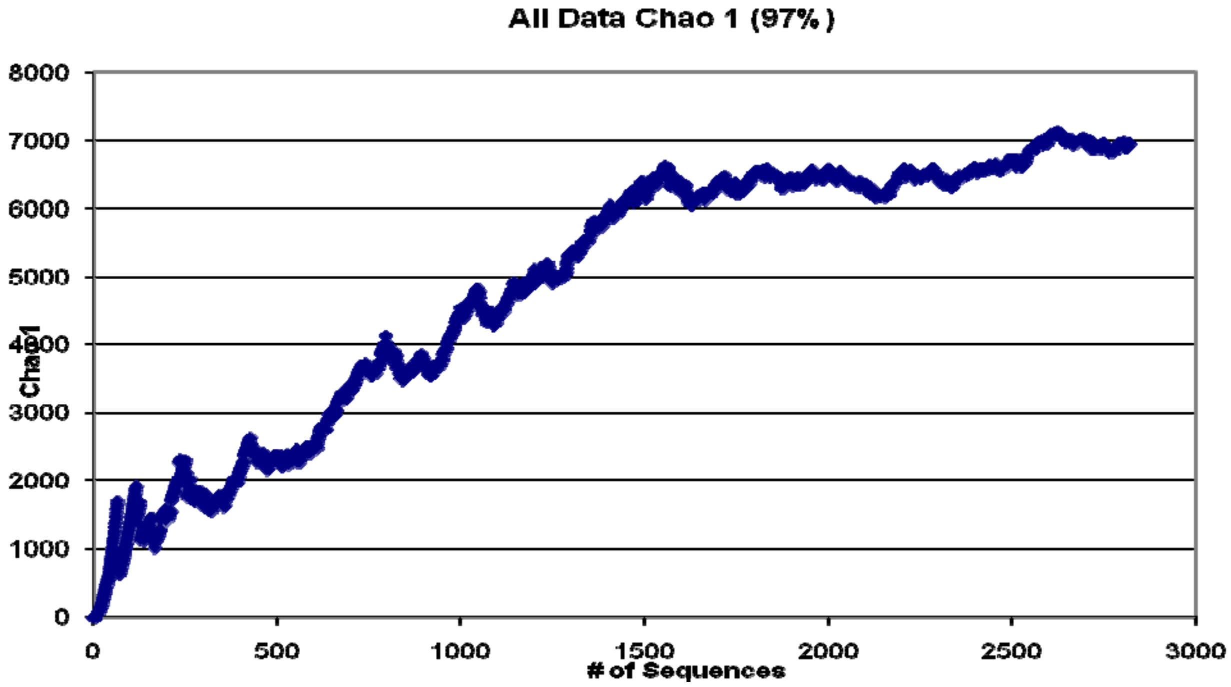


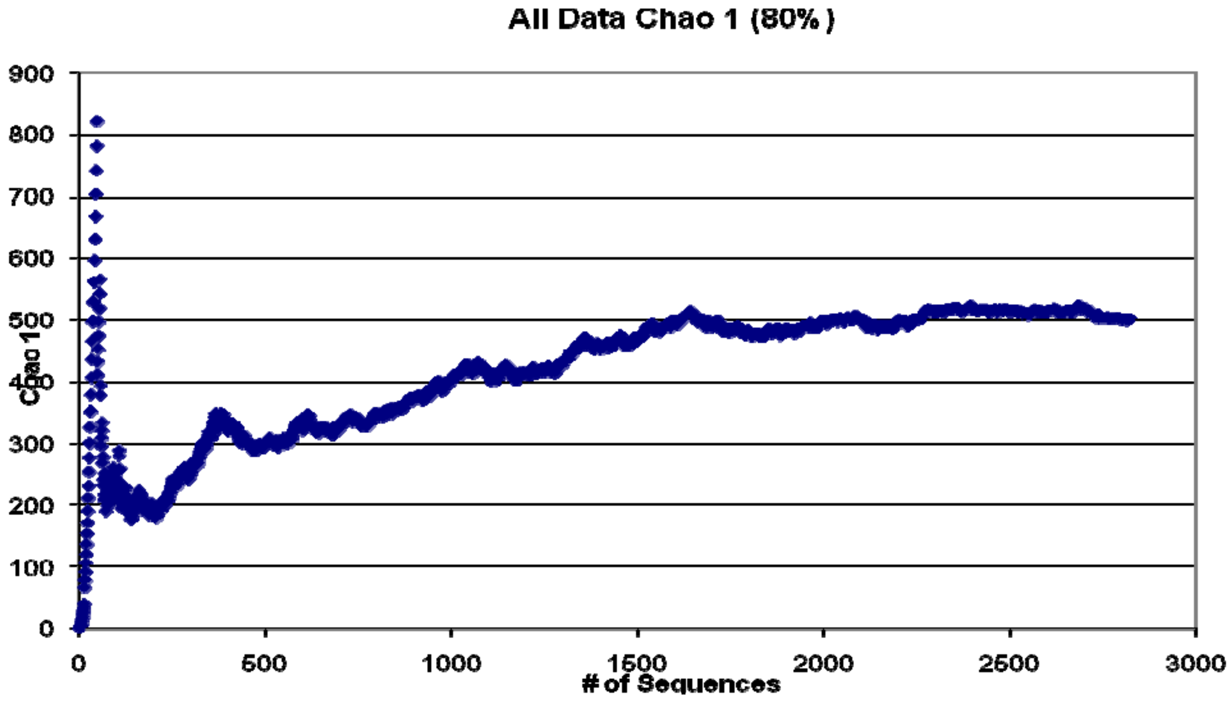
My62.L1 and EZNAGel: Lane 1-100 bp ladder; Lane 2-My62.L1.1.1; Lane 3-My62.L1.1.2; Lane 4-EZNA (other research); Lane 5: EZNA (other research); Lane 6-My62.L1.1.21; Lane 7-My62.L1.2.21; Lane 8-My62.L1.3.21; Lane 9-My62.L1.4.21; Lane 10-My62.L1.5.21

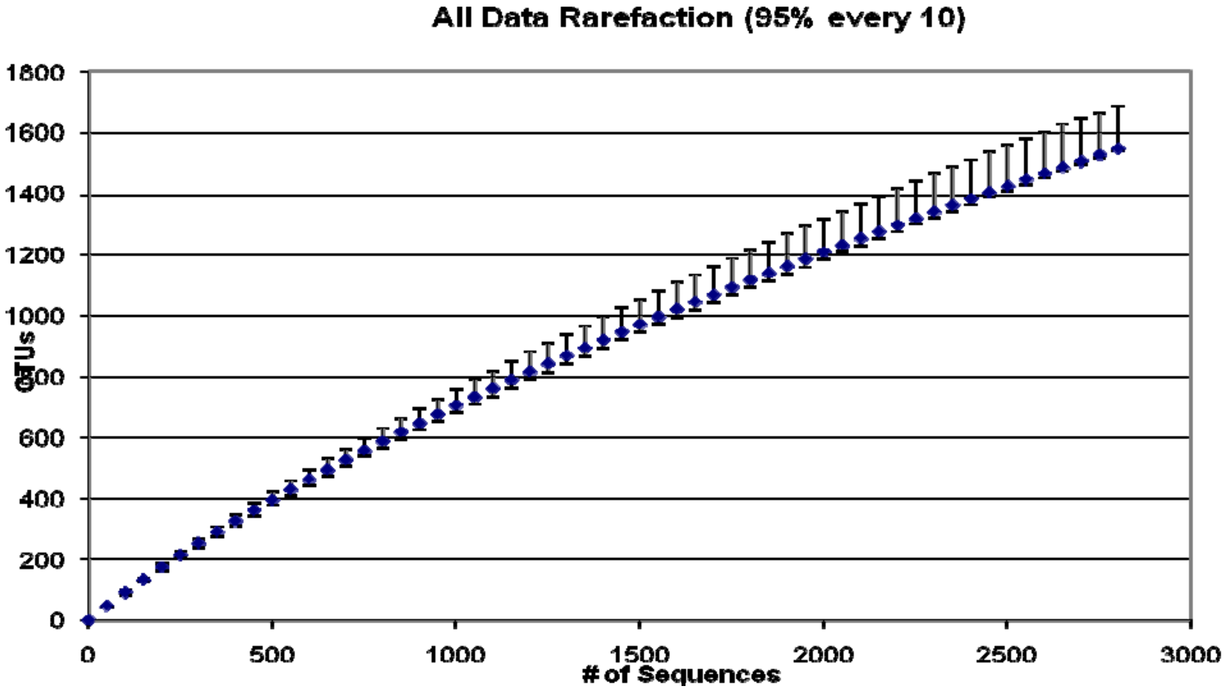
Appendix H: Richness Estimator and Rarefaction Curves

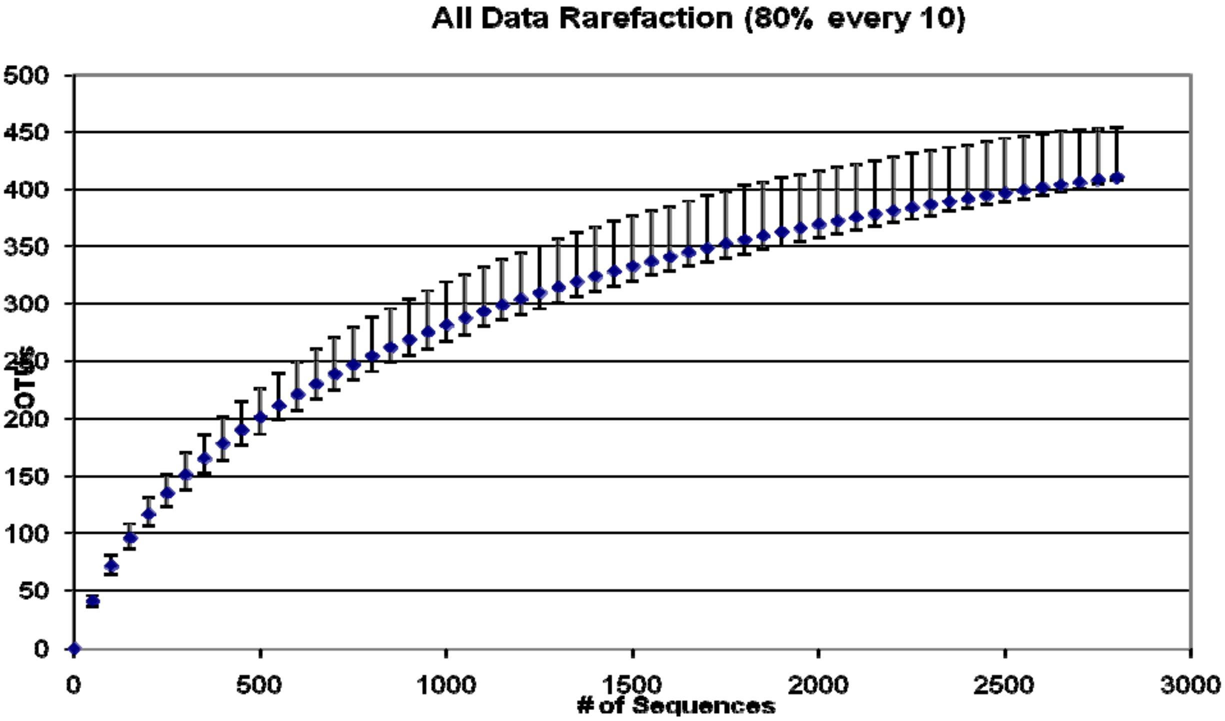


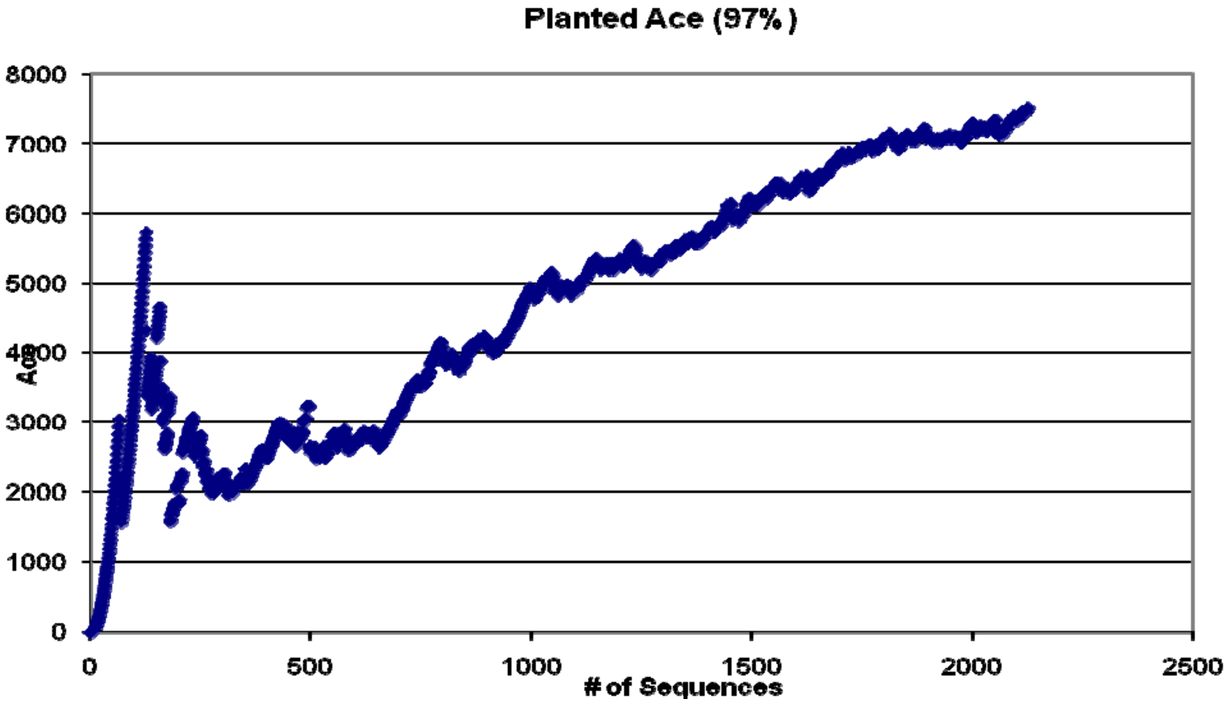


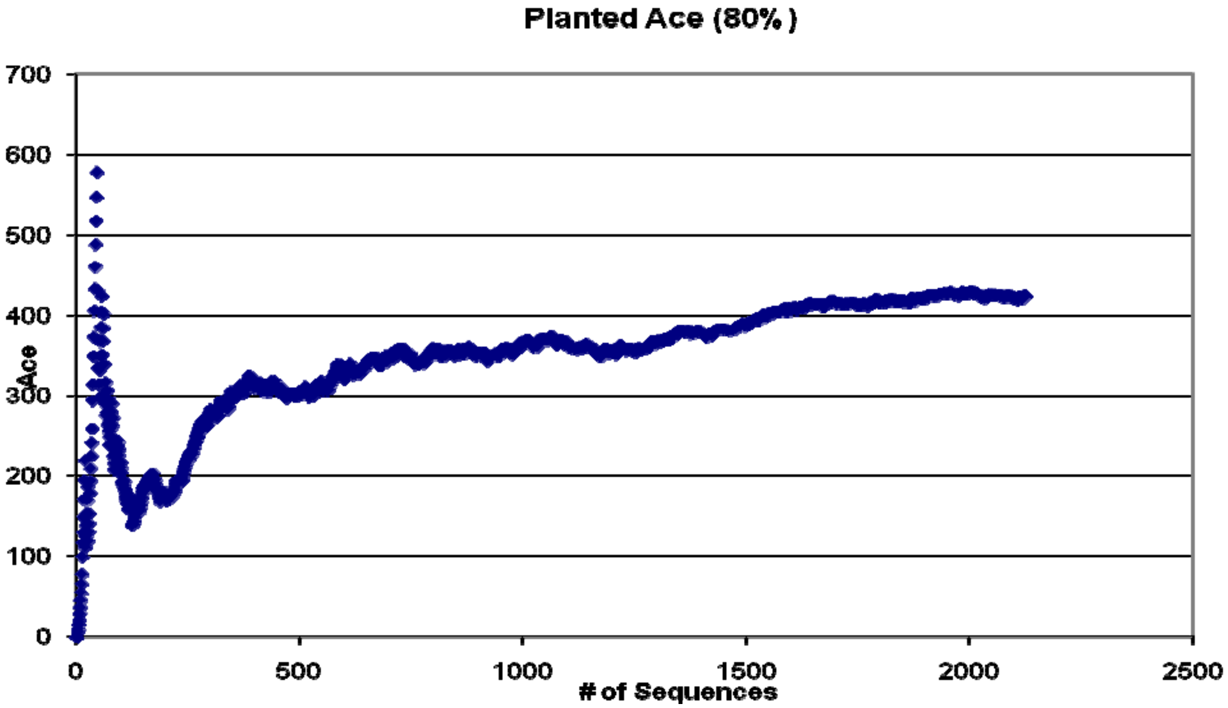

All Data Chao 1 (97%)
8000
7000
6000
5000
4000
3000
2000
1000
0
Chao1
0
500
1000
1500
2000
2500
3000
# of Sequences

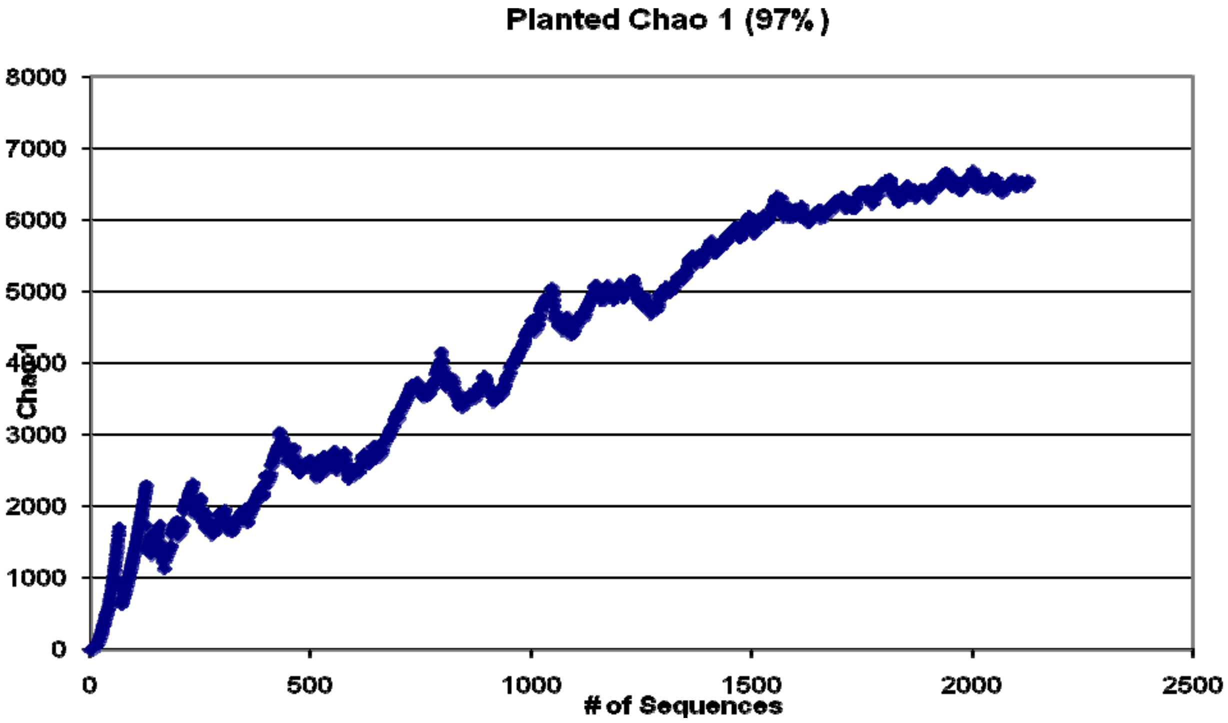

All Data Chao 1 (80%)
900
800
700
600
500
400
300
200
100
0
Chao 1
0
500
1000
1500
2000
2500
3000
# of Sequences

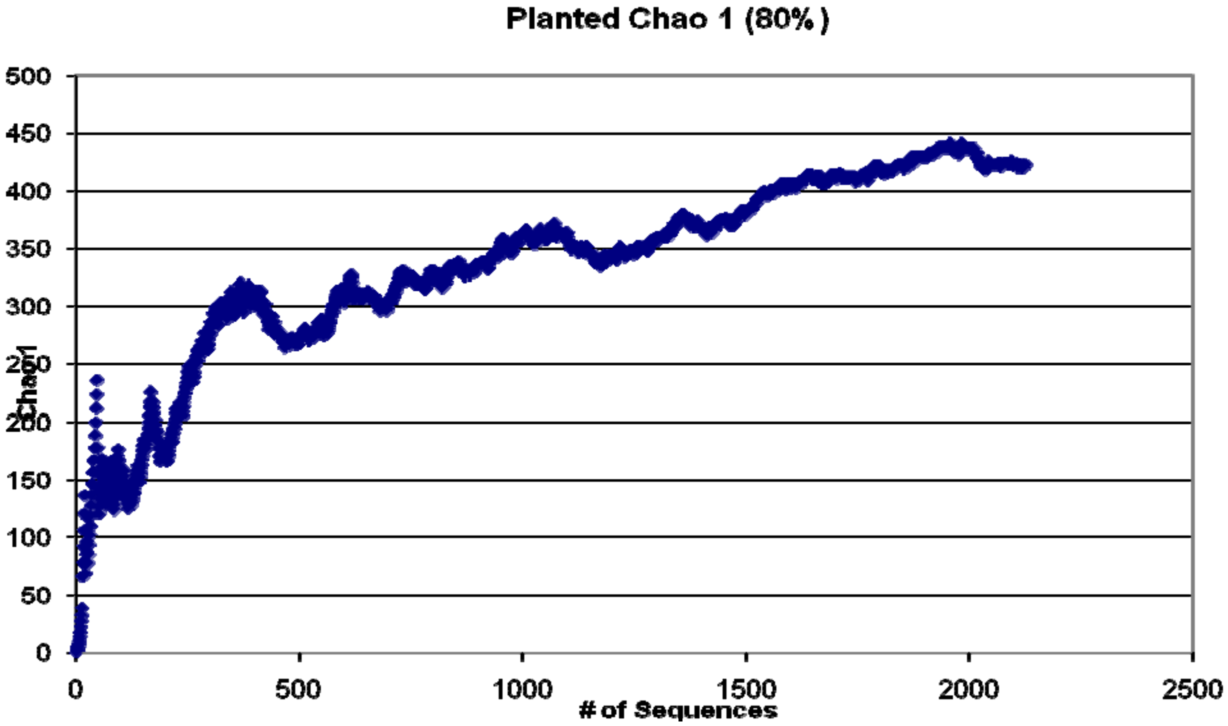

All Data Rarefaction (95% every 10)
OTUs
# of Sequences
0
200
400
600
800
1000
1200
1400
1600
1800
0
500
1000
1500
2000
2500
3000

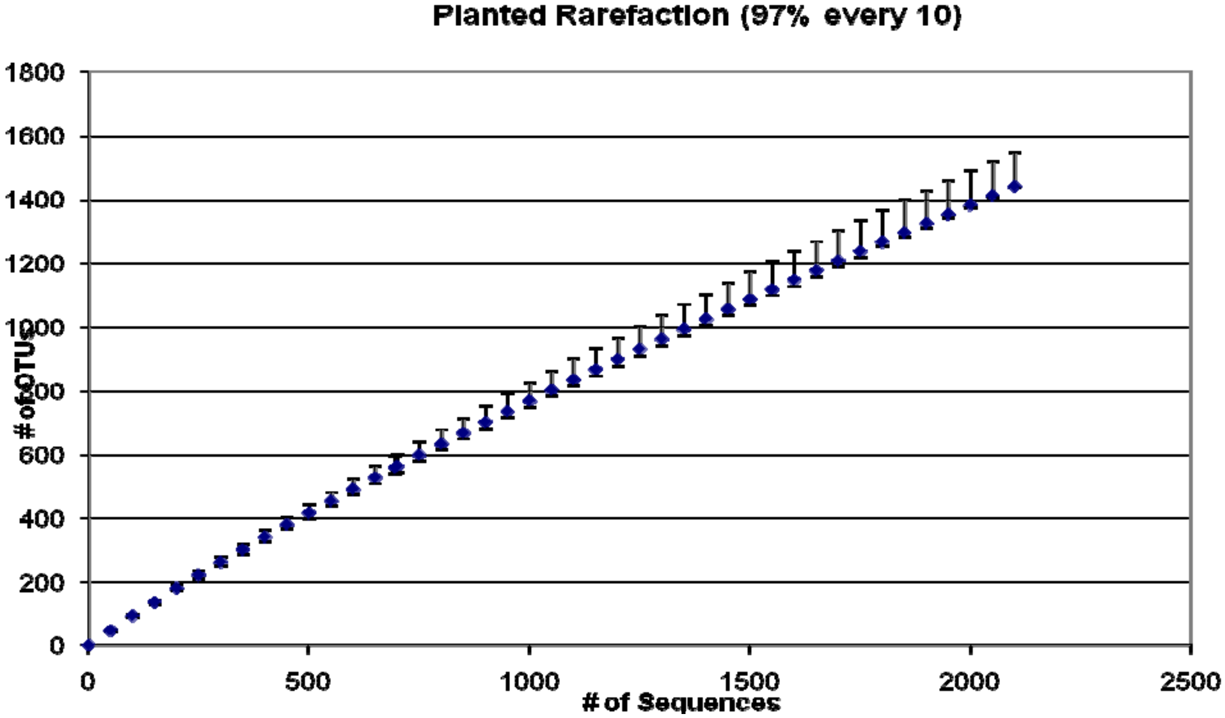

All Data Rarefaction (80% every 10)
OTUs
# of Sequences
0
50
100
150
200
250
300
350
400
450
500
0
500
1000
1500
2000
2500
3000

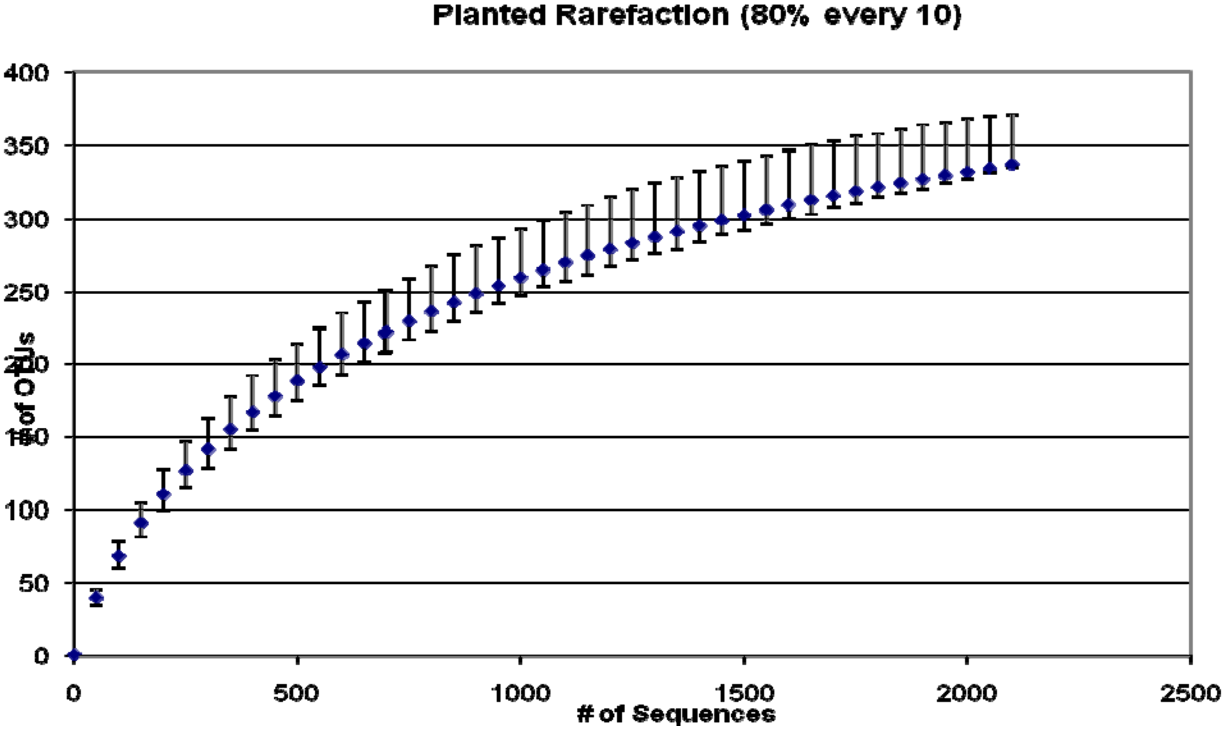

Planted Ace (97%)
Ace
8000
7000
6000
5000
4000
3000
2000
1000
0
0
500
1000
1500
2000
2500
# of Sequences


Planted Ace (80%)
Ace
700
600
500
400
300
200
100
0
0
500
1000
1500
2000
2500
# of Sequences

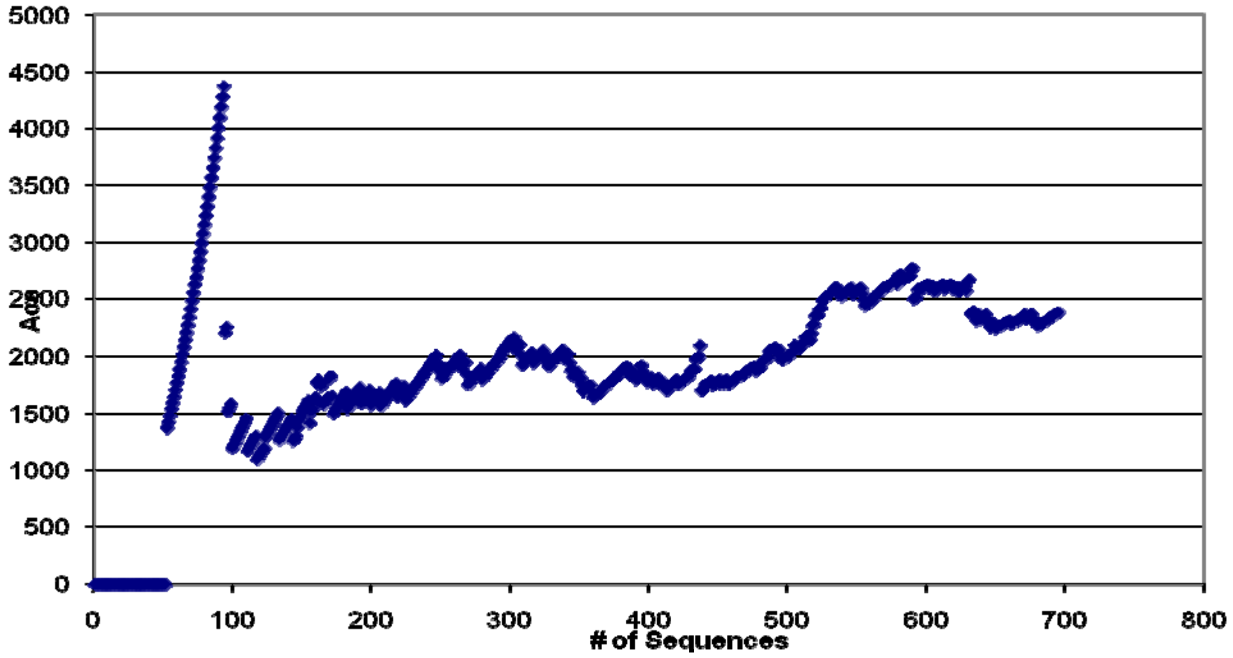

Planted Chao 1 (97%)
8000
7000
6000
5000
4000
3000
2000
1000
0
Chao1
0
500
1000
1500
2000
2500
# of Sequences


Planted Chao 1 (80%)
500
450
400
350
300
250
200
150
100
50
0
Chao1
0
500
1000
1500
2000
2500
# of Sequences

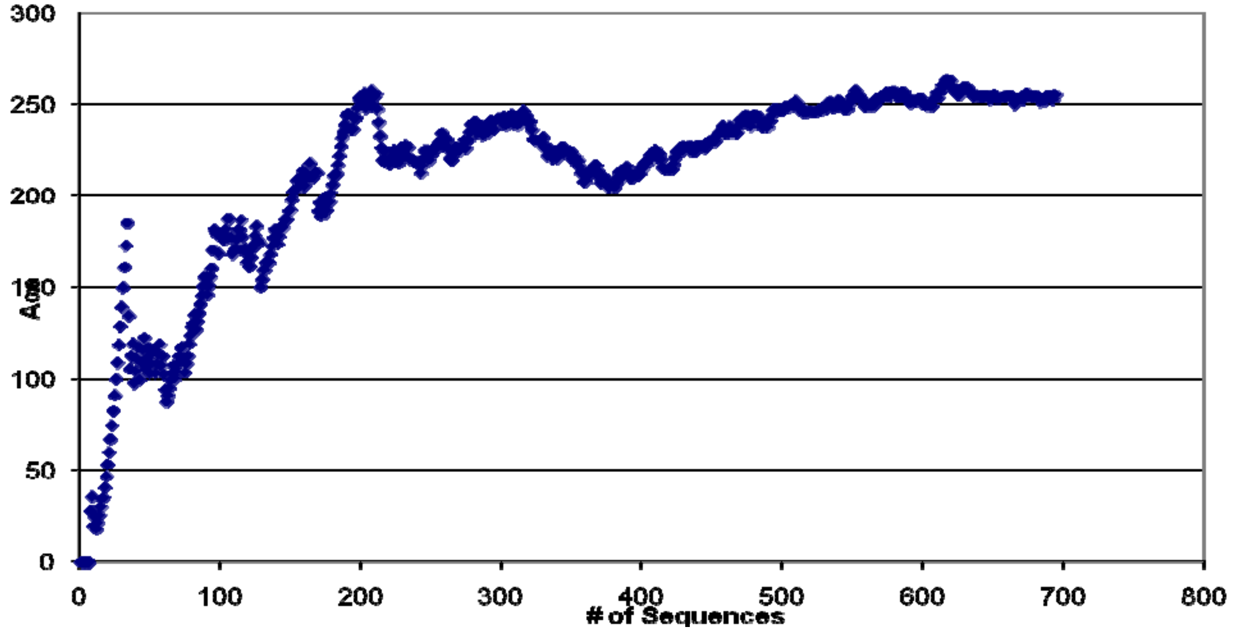

Planted Rarefaction (97% every 10)
1800
1600
1400
1200
1000
800
600
400
200
0
# of OTUs
0
500
1000
1500
2000
2500
# of Sequences

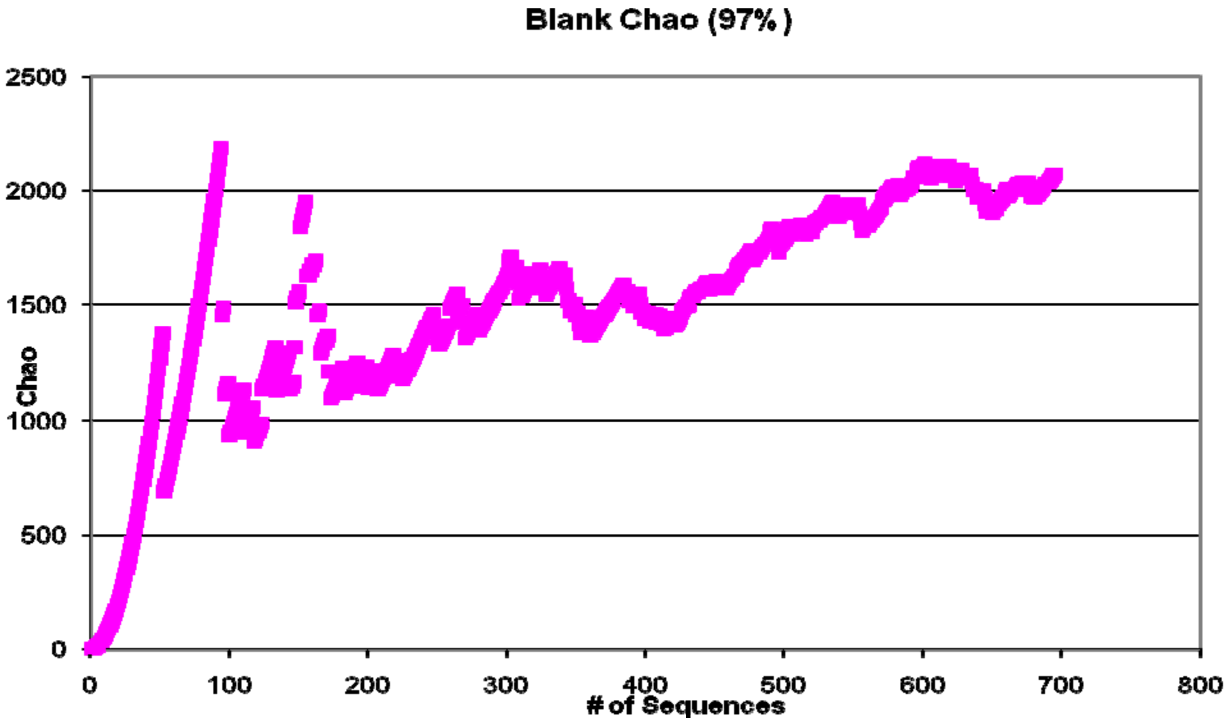

Planted Rarefaction (80% every 10)
400
350
300
250
200
150
100
50
0
# of OTUs
0
500
1000
1500
2000
2500
# of Sequences

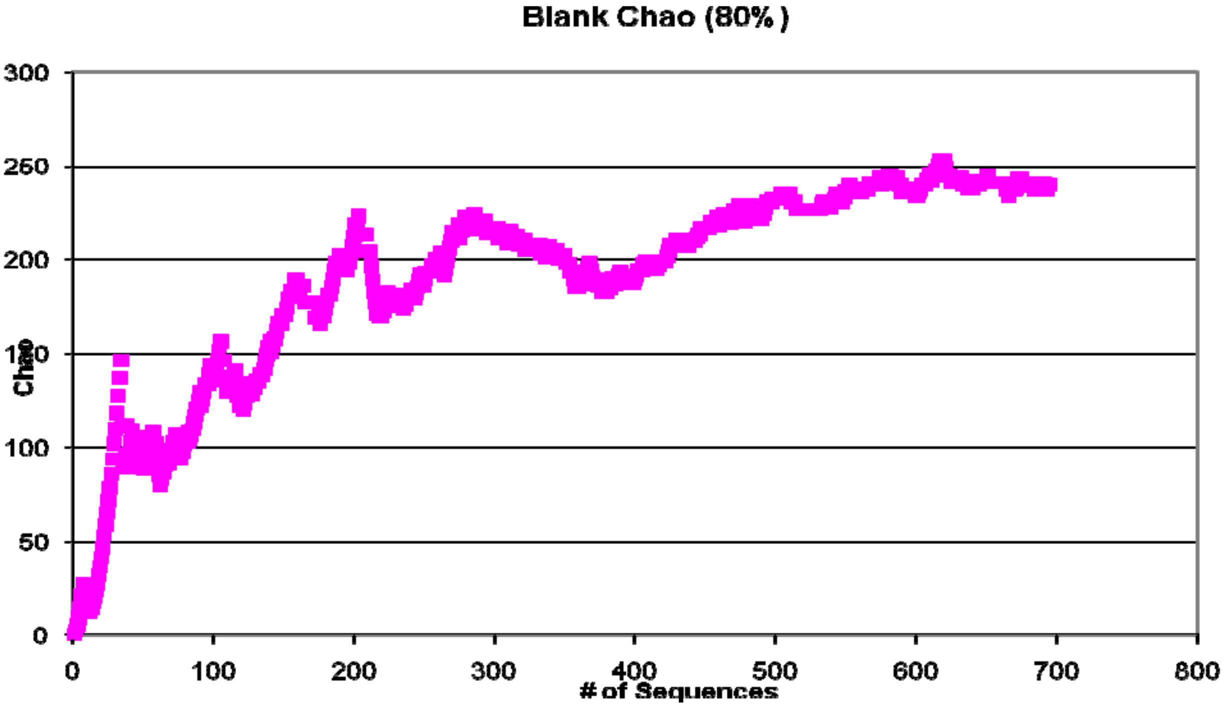
Blank Ace (97%)

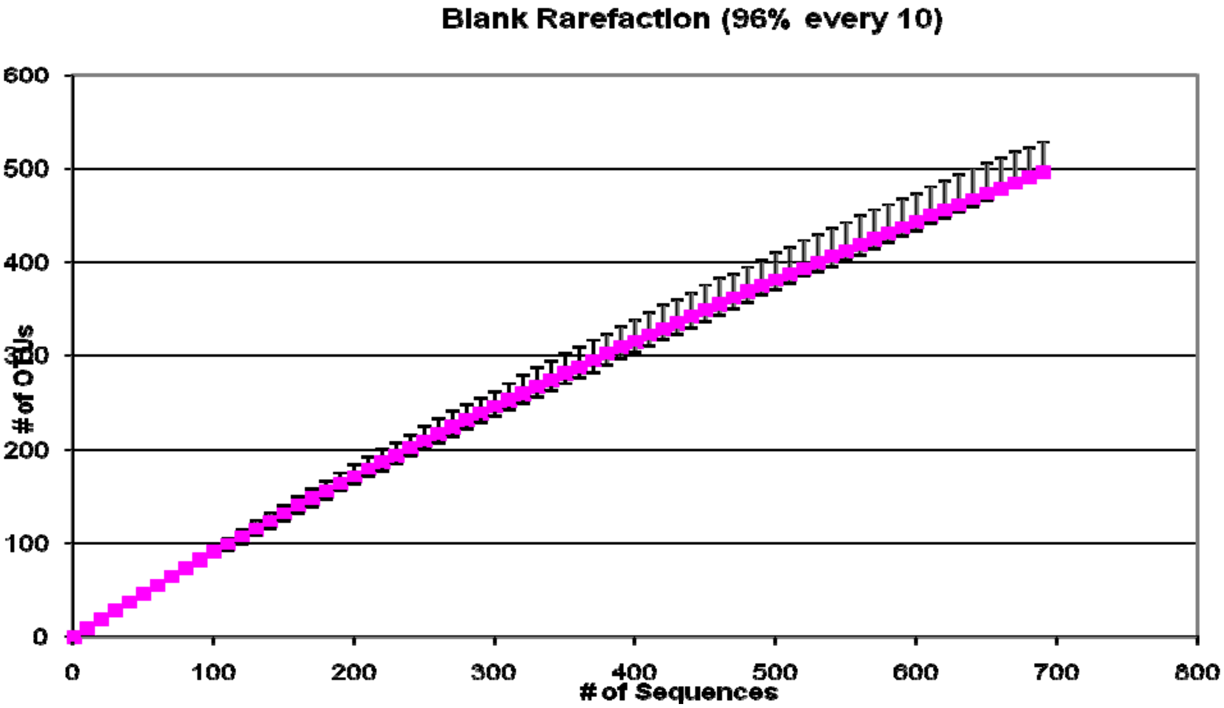


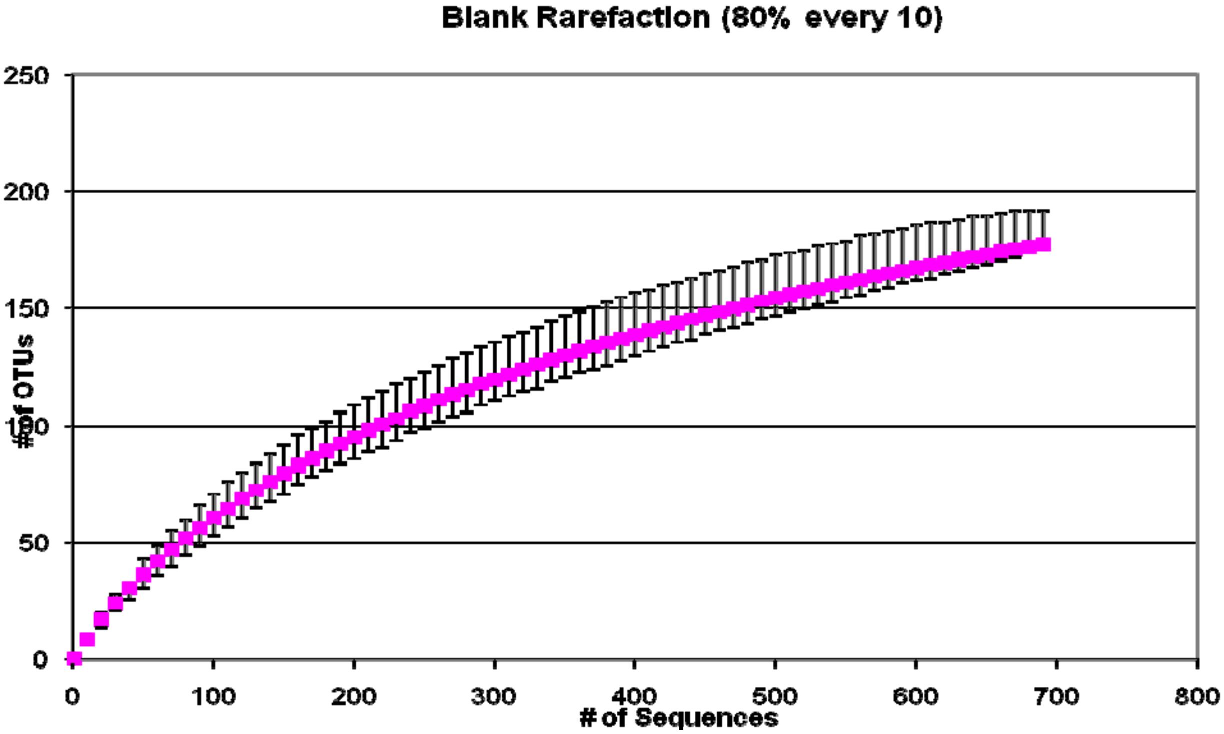
Blank Ace (80%)

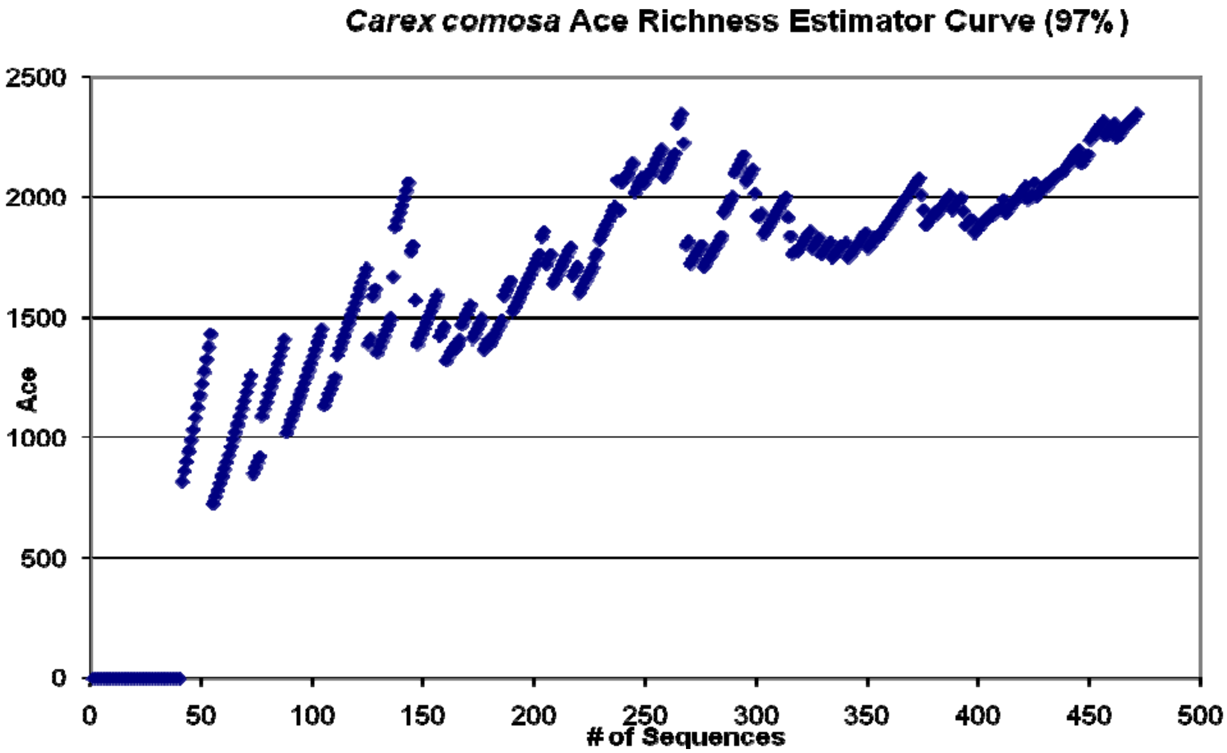


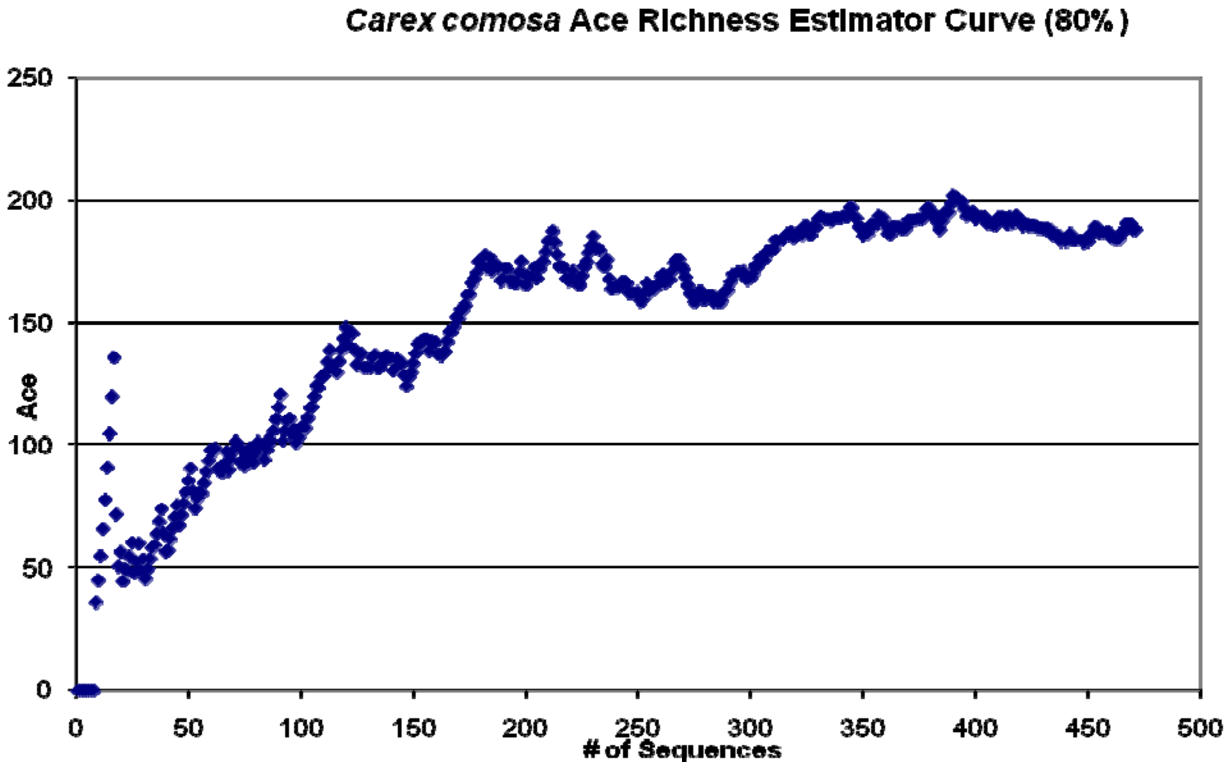

Blank Chao (97%)
2500
2000
1500
1000
500
0
Chao
0
100
200
300
400
500
600
700
800
# of Sequences

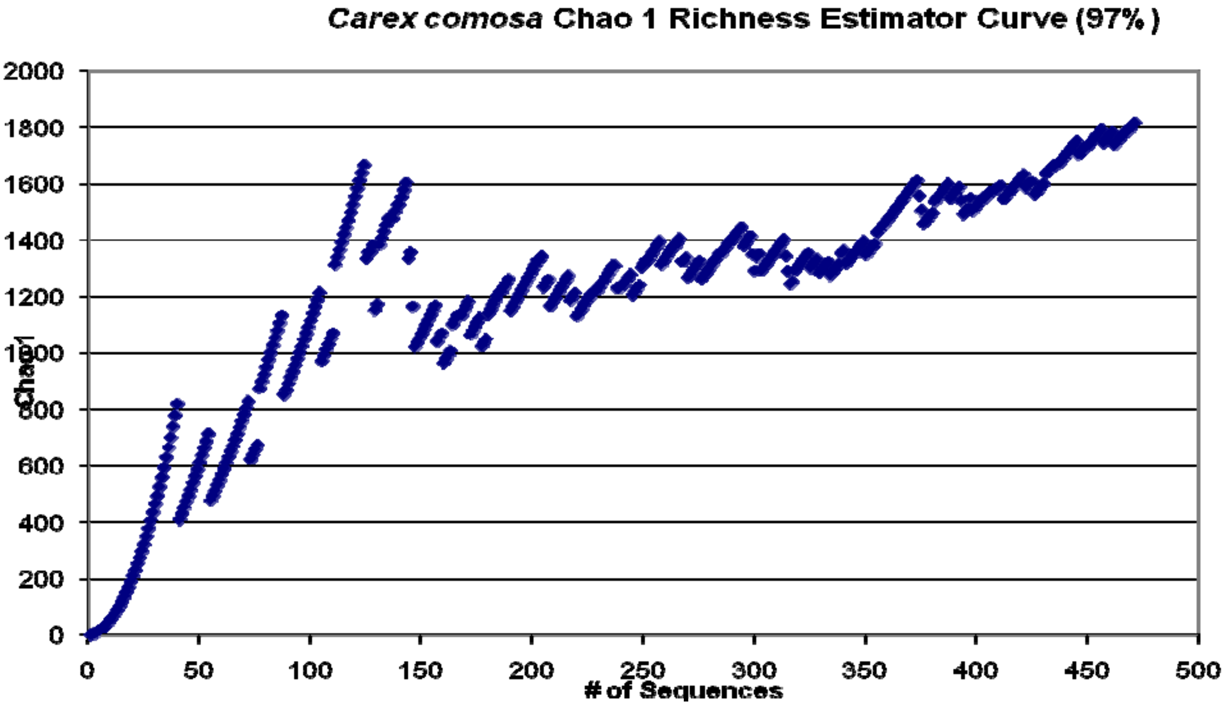

Blank Chao (80%)
300
250
200
150
100
50
0
Chao
0
100
200
300
400
500
600
700
800
# of Sequences

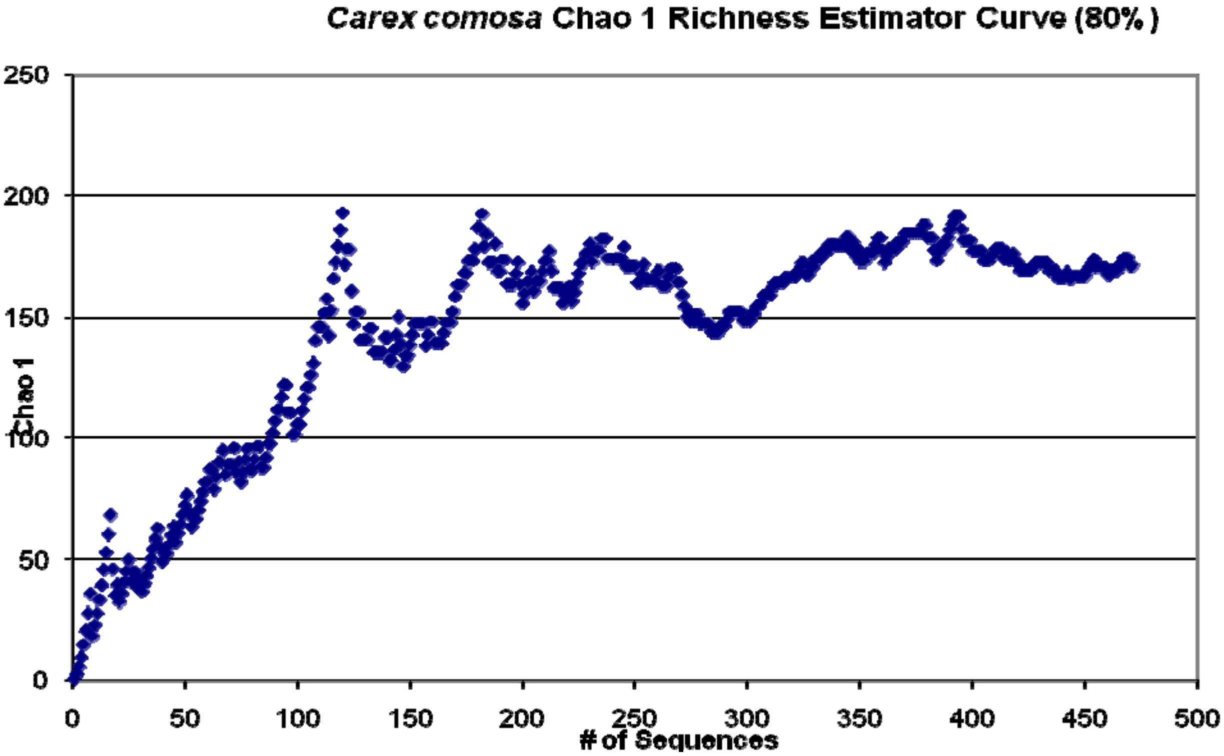

Blank Rarefaction (96% every 10)
# of OTUs
# of Sequences
0
100
200
300
400
500
600
0
100
200
300
400
500
600
700
800

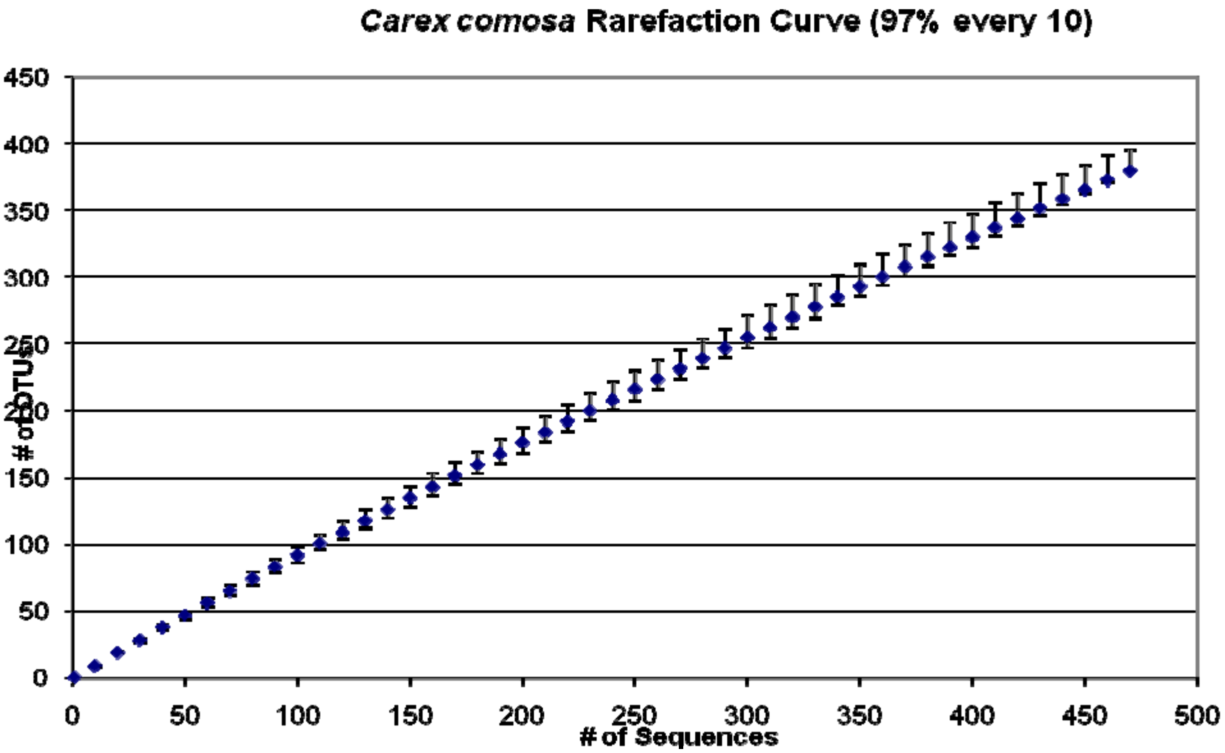

Blank Rarefaction (80% every 10)
# of OTUs
# of Sequences
0
50
100
150
200
250
0
100
200
300
400
500
600
700
800


Carex comosa Ace Richness Estimator Curve (97%)
Ace
2500
2000
1500
1000
500
0
0
50
100
150
200
250
300
350
400
450
500
# of Sequences

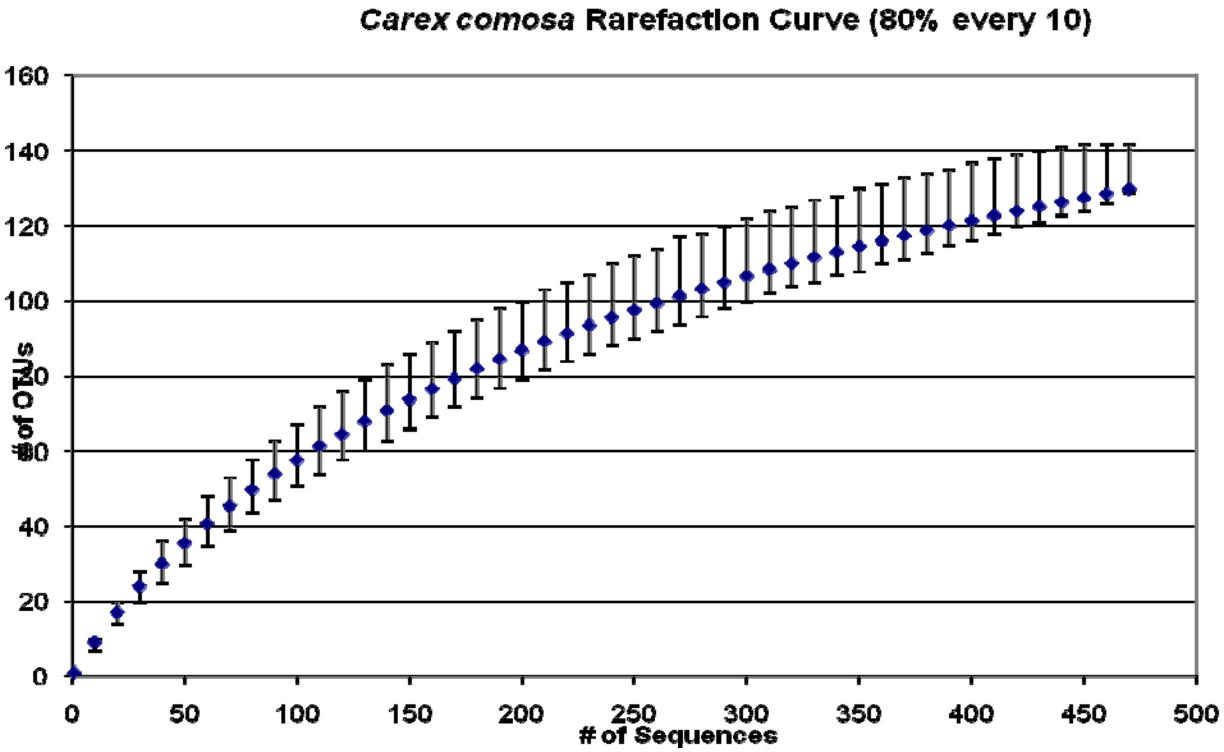

Carex comosa Ace Richness Estimator Curve (80%)
Ace
250
200
150
100
50
0
0
50
100
150
200
250
300
350
400
450
500
# of Sequences


Carex comosa Chao 1 Richness Estimator Curve (97%)
Chao 1
2000
1800
1600
1400
1200
1000
800
600
400
200
0
0
50
100
150
200
250
300
350
400
450
500
# of Sequences

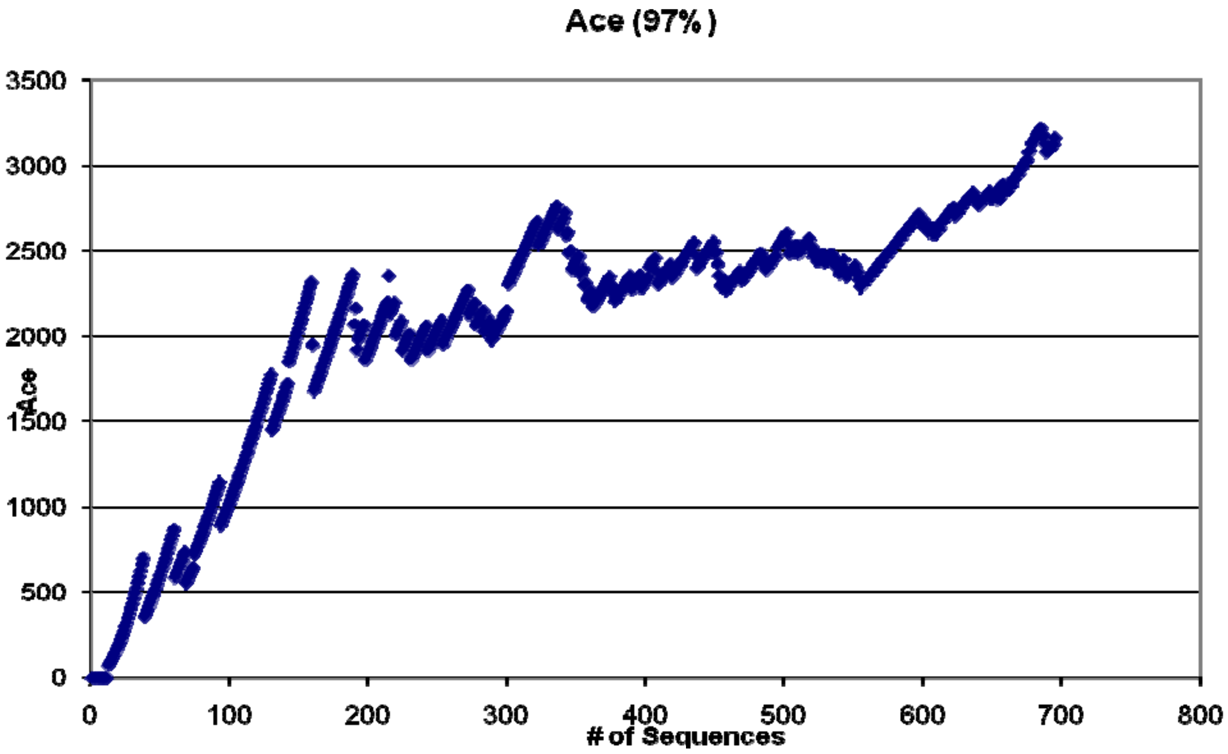

Carex comosa Chao 1 Richness Estimator Curve (80%)
Chao 1
250
200
150
100
50
0
0
50
100
150
200
250
300
350
400
450
500
# of Sequences

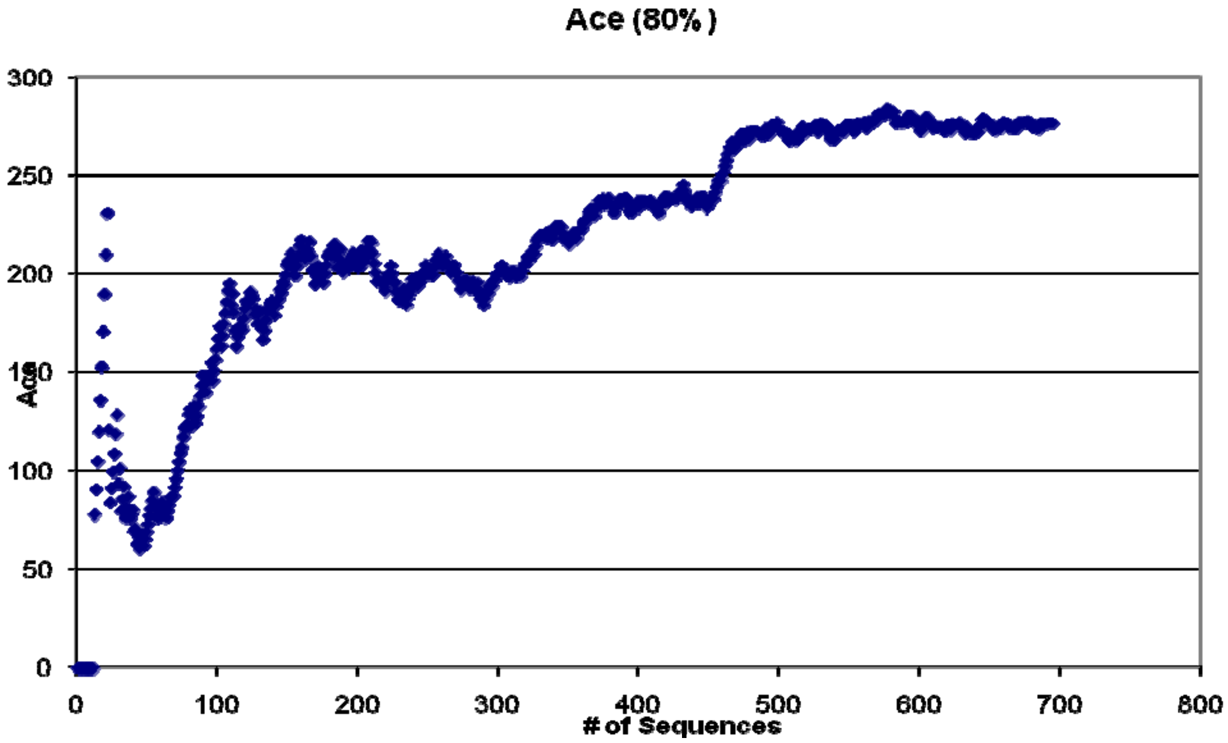


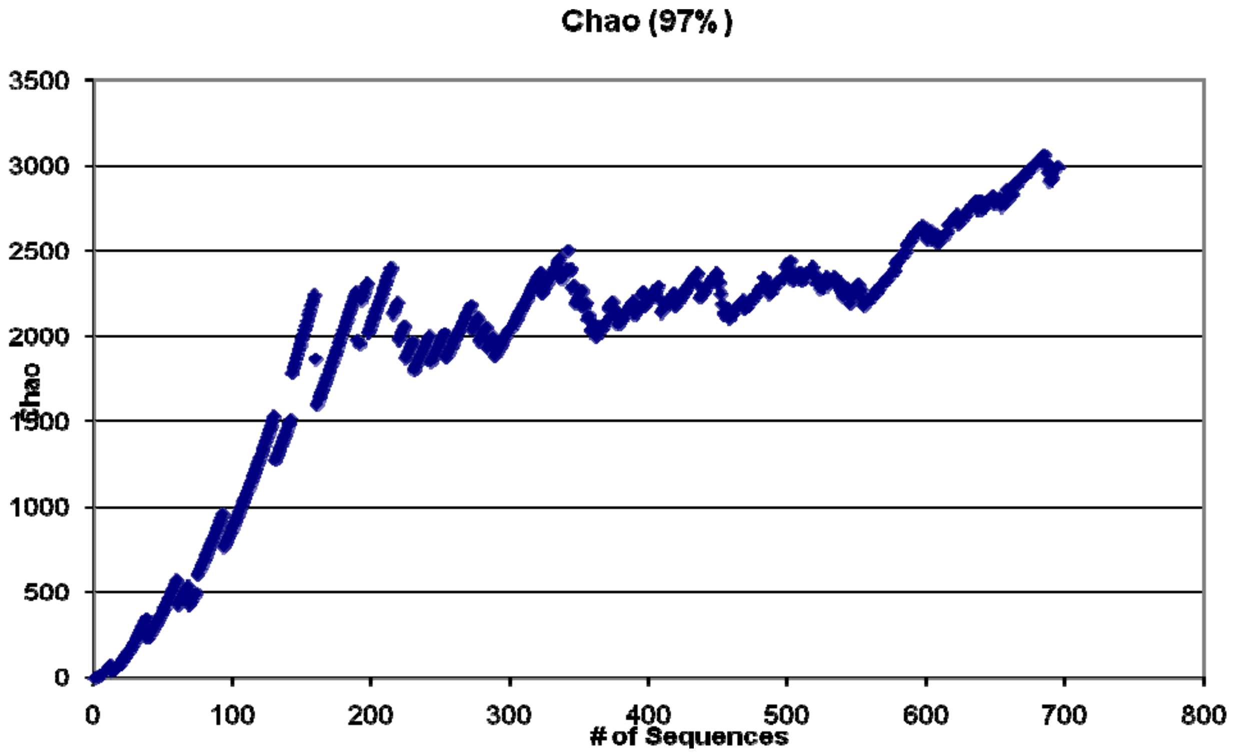
Carex comosa Rarefaction Curve (97% every 10)

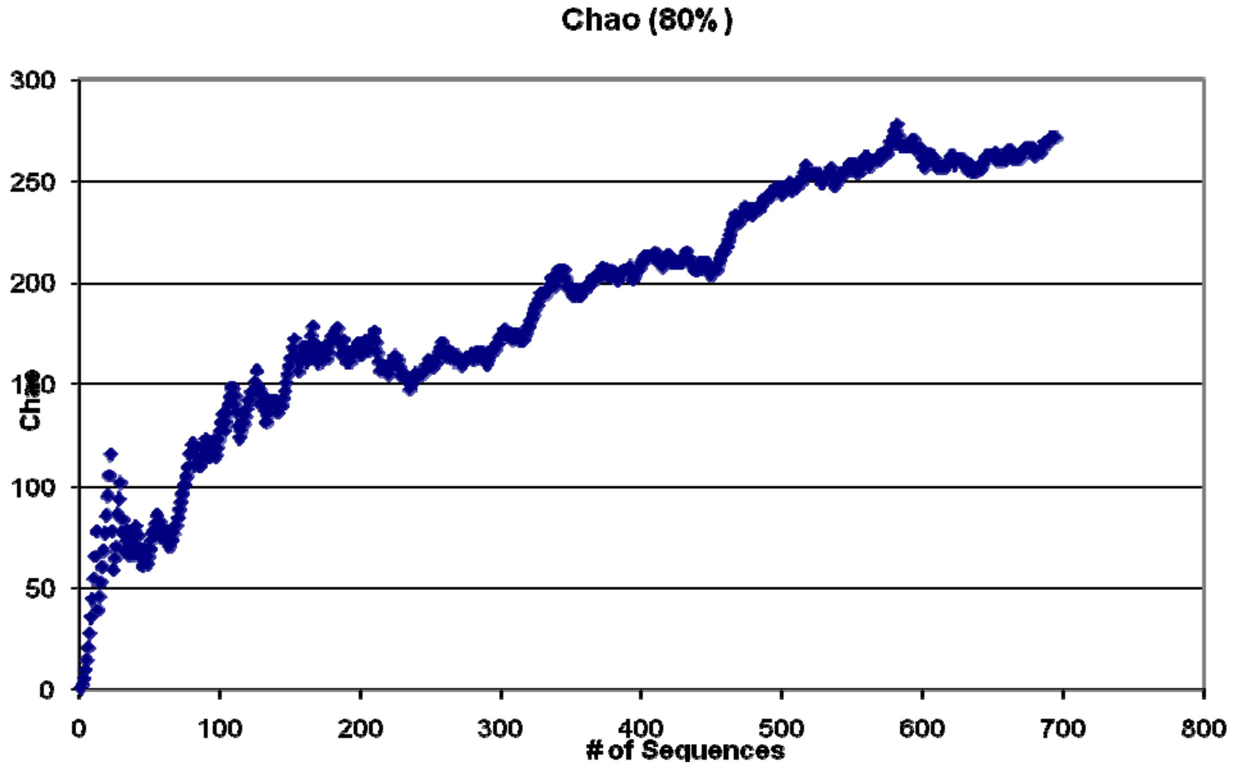


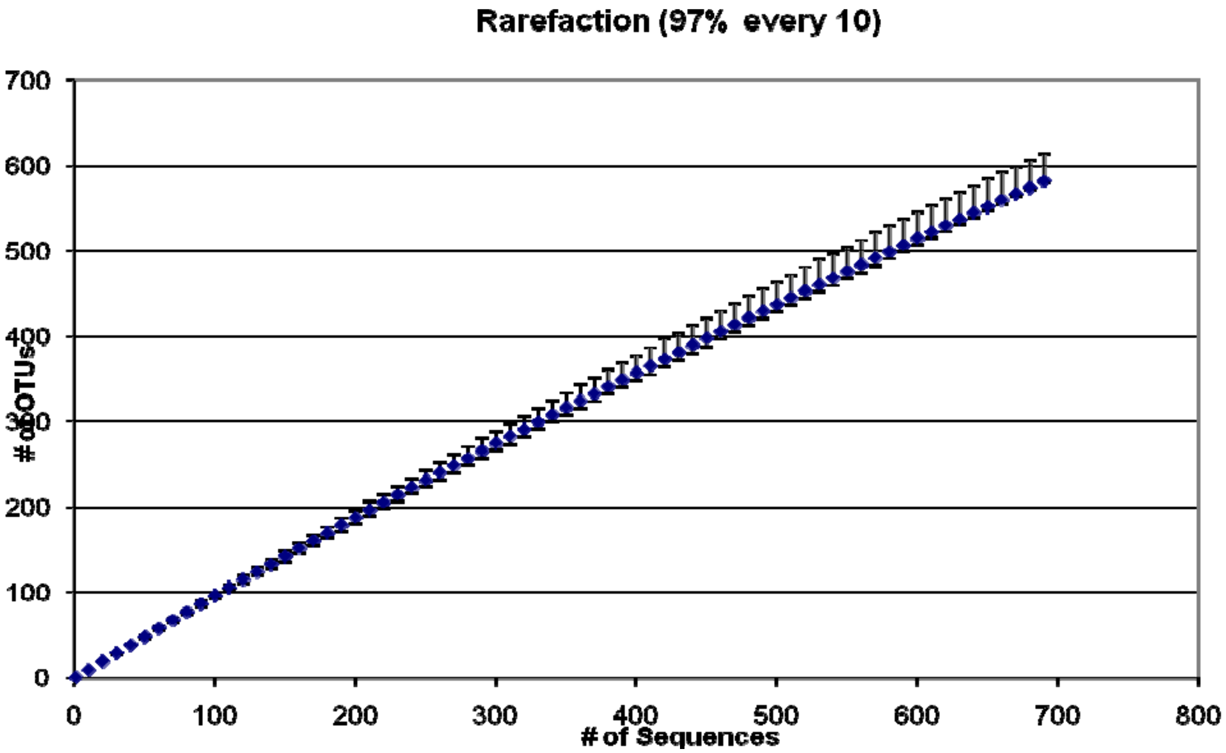
Carex comosa Rarefaction Curve (80% every 10)

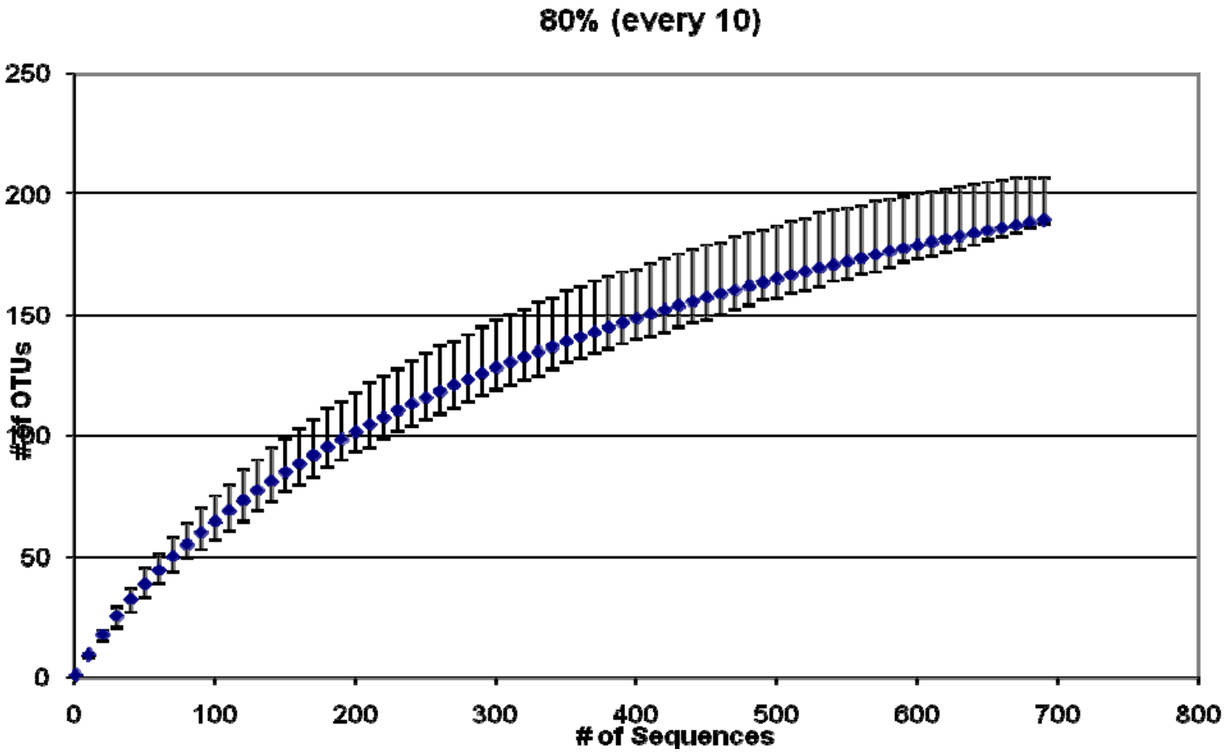

Ace (97%)
3500
3000
2500
2000
1500
1000
500
0
Ace
0
100
200
300
400
500
600
700
800
# of Sequences

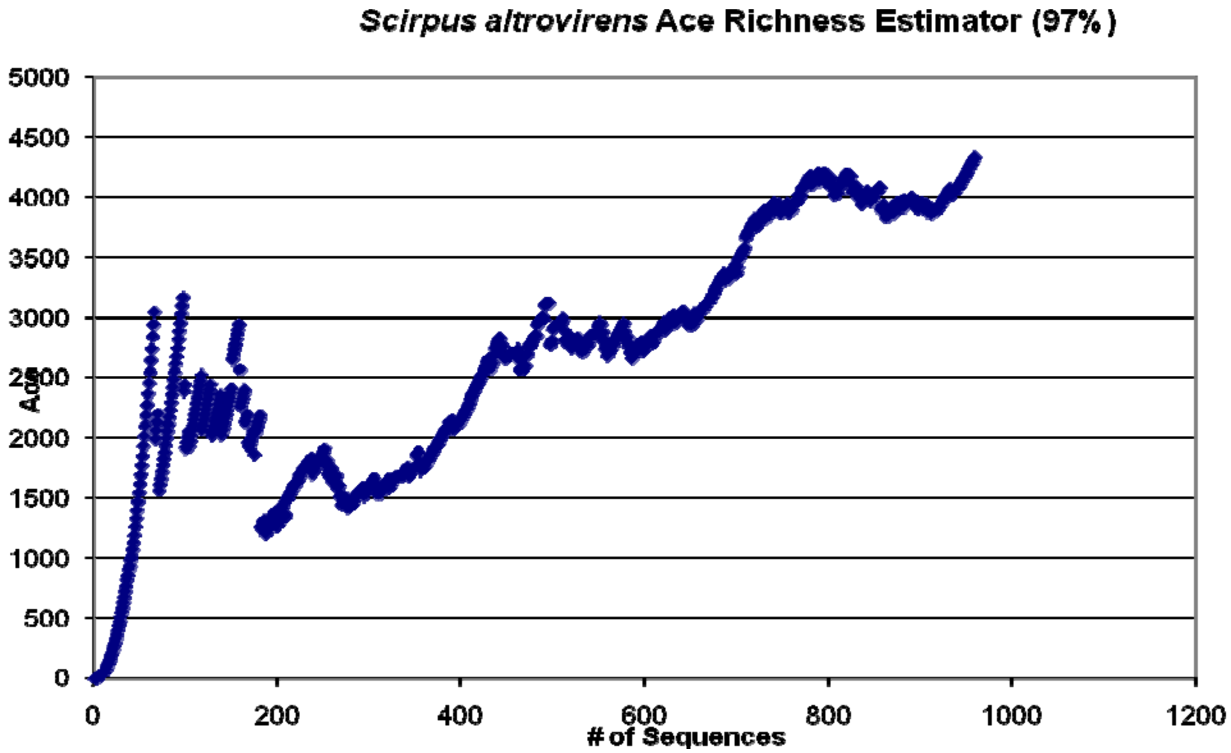

Ace (80%)
300
250
200
150
100
50
0
Ace
0
100
200
300
400
500
600
700
800
# of Sequences

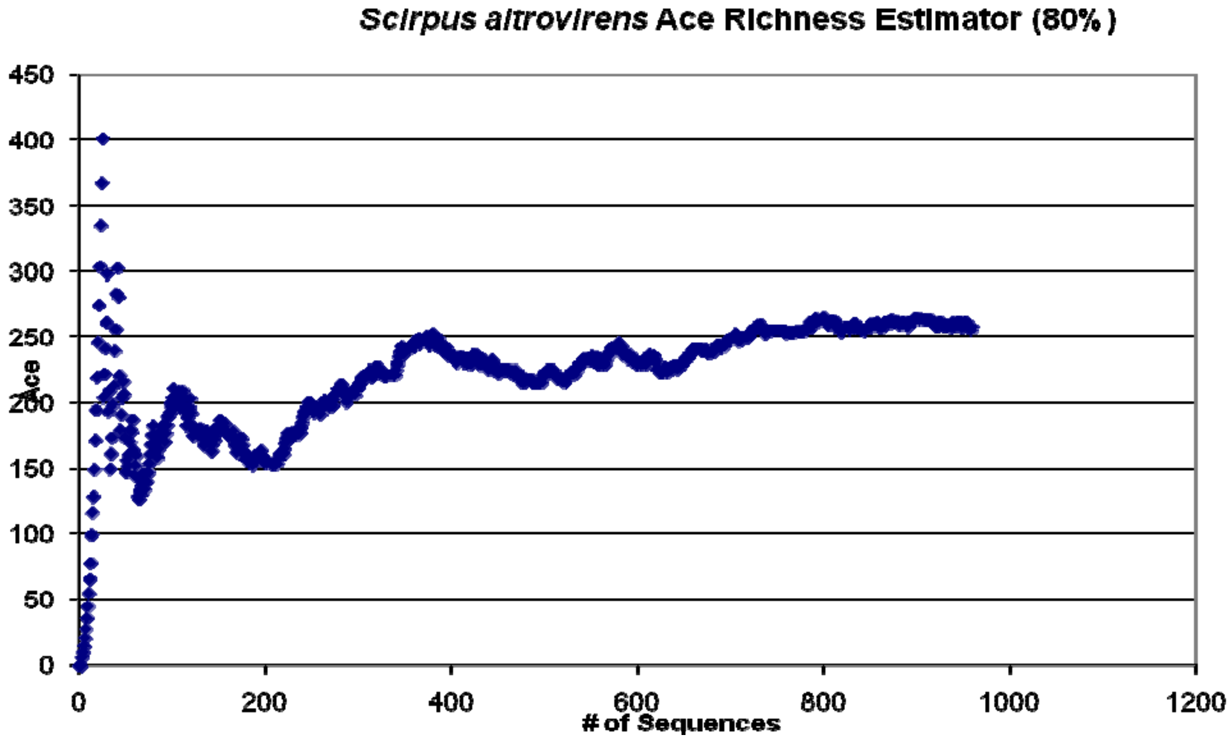

Chao (97%)
3500
3000
2500
2000
1500
1000
500
0
Chao
0
100
200
300
400
500
600
700
800
# of Sequences

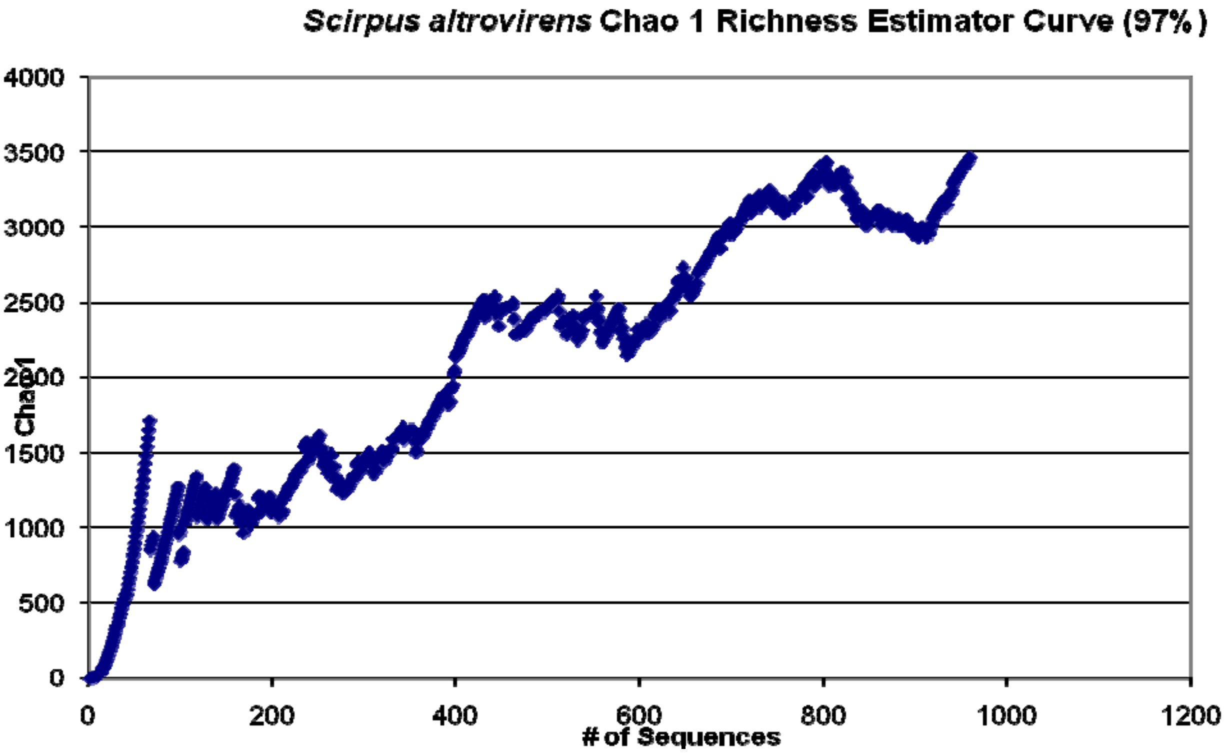

Chao (80%)
300
250
200
150
100
50
0
Chao
0
100
200
300
400
500
600
700
800
# of Sequences


Rarefaction (97% every 10)
700
600
500
400
300
200
100
0
# of OTUs
0
100
200
300
400
500
600
700
800
# of Sequences

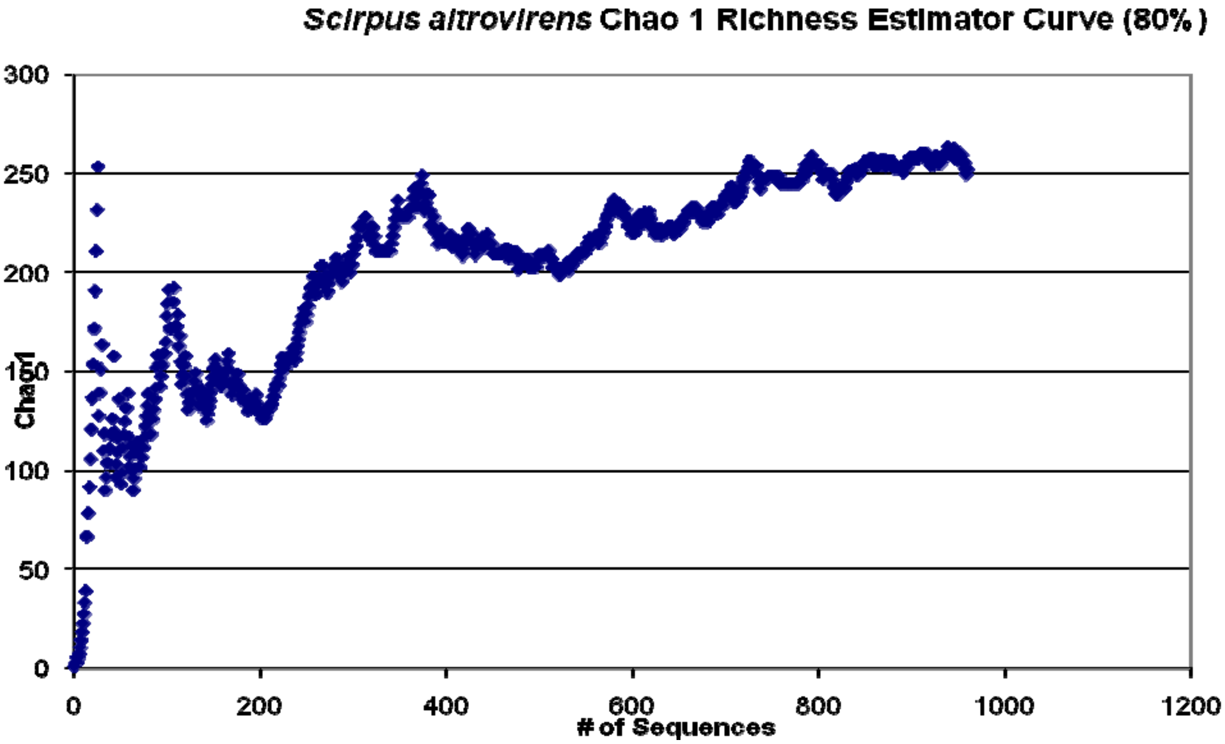

80% (every 10)
250
200
150
100
50
0
# of OTUs
0
100
200
300
400
500
600
700
800
# of Sequences


Scirpus altrovirens Ace Richness Estimator (97%)
Ace
5000
4500
4000
3500
3000
2500
2000
1500
1000
500
0
0
200
400
600
800
1000
1200
# of Sequences

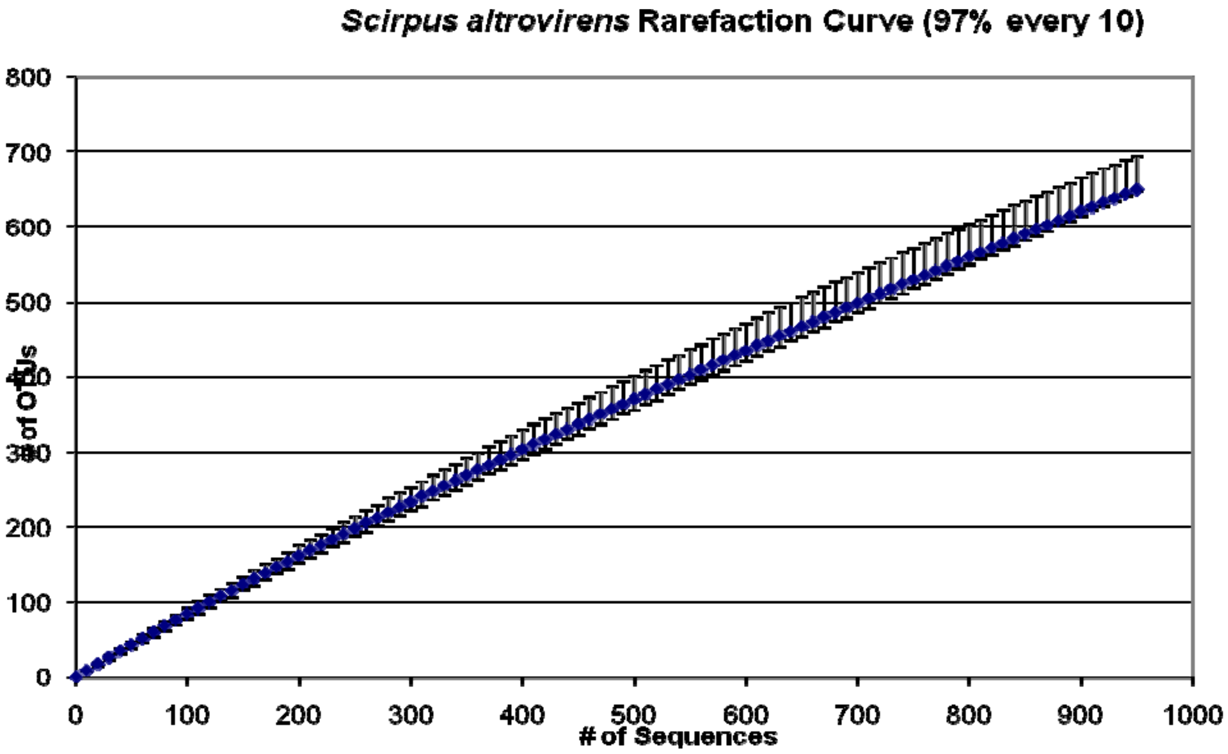

Scirpus altrovirens Ace Richness Estimator (80%)
Ace
450
400
350
300
250
200
150
100
50
0
0
200
400
600
800
1000
1200
# of Sequences

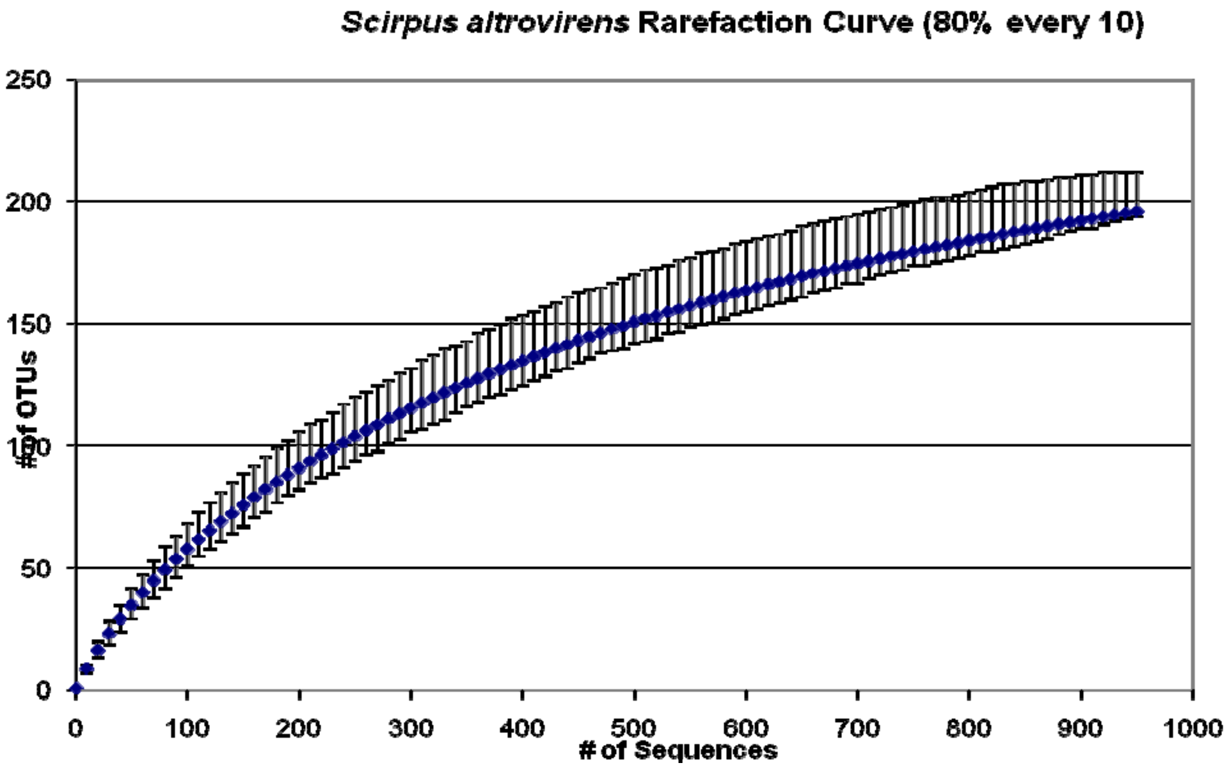


Scirpus altrovirens Chao 1 Richness Estimator Curve (97%)



Scirpus altrovirens Chao 1 Richness Estimator Curve (80%)


Scirpus altrovirens Rarefaction Curve (97% every 10)
# of OTUs
0
100
200
300
400
500
600
700
800
# of Sequences
0
100
200
300
400
500
600
700
800
900
1000


Scirpus altrovirens Rarefaction Curve (80% every 10)
# of OTUs
0
50
100
150
200
250
# of Sequences
0
100
200
300
400
500
600
700
800
900
1000

Bibliography

Altschul, Stephen F., *et al*. (1990). Basic Local Alignment Search Tool. *Journal of Molecular Biology*, 215, 403-410.

Amon, James P., *et al*. (2007). Development of a Wetland Constructed for the Treatment of Groundwater Contaminated by Chlorinated Ethenes. *Ecological Engineering*, 30, 51-66.

Baker, G.C. *et al*. (2003). Review and re-analysis of domain-specific 16S Primers. *Journal of Microbiological Methods*, 55, 541-555.

Bardgett, R.D., *et al*. (1999). Plant species and nitrogen effects on soil biological properties of temperate upland grasslands. *Functional Ecology*, 13, 650-660.

Ben-Dov, Eitan, *et al*. (2006). Advantage of using Inosine at the 3' Termini of 16S rRNA gene universal primers for the study of microbial diversity. *Applied and Environmental Microbiology*, 72(11), 6902-6906.

Bezemer, Martinjn T., *et al*. (2006). Plant species and functional group effects on abiotic and microbial soil properties and plant-soil feedback responses in two grasslands. *Journal of Ecology*, 94, 893-904.

Bik, Elisabeth M., *et al*. (2006). Molecular analysis of the bacterial microbiota in the human stomach. PNAS, 103(3), 732-737.

Bishop, Ethan C. (2006). "Molecular Characterization of Wetland Soil Bacterial Community In Constructed Mesocosms." Dept of Systems and Engineering Management (M.S. ed). Wright-Patterson AFB, OH: Air Force Institute of Technology.

Bleckmann, Charles. (2007). Class handout, ENVR 772, Subsurface Restoration. School of Engineering and Management, Air Force Institute of Technology, Wright-Patterson Air Force Base, OH.

Bond, P.L., P. Hugenholtz, J. Keller, and L.L. Blackall. (1995). Bacterial community structures of phosphate-removing and non-phosphate-removing activated sludges from sequencing batch reactors. *Applied Environmental Microbiology*, 61, 1910-1916.

Borneman, James., *et al*. (1996). Molecular microbial diversity of an agricultural soil in Wisconsin. *Applied and Environmental Microbiology*, 62(6), 1935-1943.

Chaffin, D.O. & Rubens, C.E. (1998). Blue/White screening of recombinant plasmids in gram-positive bacteria by interruption of alkaline phosphatase gene (*pho*Z) expression. *Gene*, 219, 91-99.

Chao, A. (1984). Nonparametric estimation of the number of classes in a population. *Scandinavian Journal of Statistics*, 11, 265-270.

Cho, Jang-Cheon, *et al*. (2004). *Lentisphaera araneosa* gen. nov., sp. Nov, a transparent exopolymer producing marine bacterium, and the description of a novel bacterial phylum, *Lentisphaerae*. *Environmental Microbiology*, 6(6), 611-621.

Clarke,K.R. (1993). Non-parametric Multivariate Analyses of Changes in Community Structure. Aust J Ecology, 18, 117-143.

Clarridge, J.E., 3rd. (2004). Impact of 16S rRNA gene sequence analysis for identification of bacteria on clinical microbiology and infectious diseases. *Clinical Microbiology Reviews*, 17(4), 840-862, table of contents.

Cole, J. R., B. Chai, R. J. Farris, Q. Wang, A. S. Kulam-Syed-Mohideen, D. M. McGarrell, A. M. Bandela, E. Cardenas, G. M. Garrity, and J. M. Tiedje. (2007). The ribosomal database project (RDP-II): introducing *myRDP* space and quality controlled public data. *Nucleic Acids Research,* 35 (Database issue): D169-D172; doi: 10.1093/nar/gkl889

Cole, J.R., *et al*. (2003). The Ribosomal Database Project (RDP-II): Previewing a New Autoaligner That Allows Regular Updates and the New Prokaryotic Taxonomy. *Nucleic Acids Research*, 31(1), 442-443.

Cumming, Geoff, *et al*. (2007). Error bars in experimental biology. *Journal of Cell Biology*, 177(1), 7-11.

Difco Laboratories. (1998). *Difco Manual*. Division of Becton Dickinson and Company. Eleventh edition. Sparks, MD.

DOTUR Manual. (2005). DOTUR: Distance Based OTU and Richness Determination.

Everett, K.D.E., Bush, R.M., and Andersen, A.A. (1999). Emended description of the order *Chlamydiales*, proposal of *Parachlamydiaceae* fam. nov. and *Simkaniaceae* fam. nov., each containing one monotypic genus, revised taxonomy of the family *Chlamydiaceae*, including a new genus and five new species, and standards for the identification of organisms. *International Journal of Systematic Bacteriology*, 49, 415-440.

Felsenstein, J. (2005). PHYLIP (Phylogeny Inference Package) version 3.6. Distributed by the author. Department of Genome Sciences, University of Washington, Seattle.

Fields, J.A. & Sierra-Alvarez R. (2004). Biodegradability of chlorinated solvents and related chlorinated aliphatic compounds. *Science Dossier*, Publication 8.

Francis, Christopher A., *et al*. (2005). Ubiquity and diversity of ammonia-oxidizing archea in water columns and sediments in the ocean. *Proceeding of the National Academy of Science of the United States of America*, 102(41), 14683-14688.

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

Grayston, Susan J., *et al*. (1998). Selective influence of plant species on microbial diversity in the rhizosphere. *Soil Biology & Biochemistry*, 30(3), 369-378.

Hugenholtz, Philip., *et al*. (1998). Impact of Culture-Independent Studies on the Emerging Phylogenetic View of Bacterial Diversity. Journal of Bacteriology, 180(18), 4765-4774.

Hugenholtz, Philip., *et al*. (2001). Investigation of Candidate Division TM7, a Recently Recognized Major Lineage of the Domain *Bacteria* with No Known Pure-Culture Representatives. *Applied and Environmental Microbiology*, 67(1), 411-419.

Hughes, Jennifer *et al.* (2001). Counting the Uncountable: Statistical Approaches to Estimating Microbial Diversity. *Applied and Environmental Microbiology*, 67, 4399-4406.

Harris, Kirk J., *et al*. (2004). New Perspective on Uncultured Bacterial Phylogenetic Division OP11. *Applied Environmental Microbiology*, 70(2), 845-849.

Isenhouer, Gwyn. (2007). PhD candidate, Wright State University, Dayton, OH. Personal Correspondence.

Jannsen, Peter H. (2006). Identifying the dominant soil bacterial taxa in libraries of 16S rRNA and 16S rRNA genes. *Applied and Environmental Microbiology*, 72(3), 1719-1728.

Kadlec, Robert H. & Knight, Robert L. (1996). Treatment Wetlands. New York: Lewis Publishers.

Kemp, Paul F. & Aller, Josephine Y. (2004). Bacterial diversity in aquatic and other environments: what 16S rDNA libraries can tell us. *FEMS Microbiology Ecology*, 47, 161-177.

Kennedy, A.C. & Smith, K.L. (1995). Soil microbial diversity and the sustainability of agricultural soils. Plant and Soil, 170, 75-86.

Kowalchuck, GA., *et al*. (2000). Changes in the community structure of ammonia-oxidizing bacteria during secondary succession of calcareous grasslands. *Environmental Microbiology*, 2, 99-110.

Kowalchuck, George., *et al*. (2002). Effects of above-ground plant species composition and diversity on the diversity of soil-borne microorganism. *Antonie van Leeuwenhoek*, 81, 509-520.

Liu, Wen-Tso, *et al*. (1997). Characterization of microbial diversity by determining terminal restriction fragment length polymorphisms of genes encoding 16S rRNA. *Applied and Environmental Microbiology*, 63(11), 4516-4522.

Marrs, R.H., *et al*. (1991). Soil chemistry and leaching losses of nutrients from semi-natural grassland and arable soils on three contrasting parent materials. *Biological Conservation*, 57, 257-271.

Marschner, P., *et al*. (2001). Soil and plant specific effects on bacterial community composition in the rhizosphere. *Soil Biology & Biochemistry*, 33, 1437-1445.

McClave, James T., *et al*. (2008). Statistics for Business and Economics. Upper Saddle River, NJ: Pearson Prentice Hall.

McGarvey, Jeffrey A., *et al*. (2004). Identification of bacterial populations in dairy wastewater by use of 16S rRNA gene sequences and other genetic markers. *Applied and Environmental Microbiology*, 70(7), 4267-4275.

McGlynn *et al*., (2002). A Phylogenetic analysis identifies heterogeneity among hepatocellular carcinomas. *Hepatology*, 36(6), 1341-1348.

Meithling, R., *et al*. (2000). Variation of microbial rhizosphere communities in response to crop species, soil origin, and inoculations with *Sinorhizobium meliloti* L 33. *Microbial Ecology*, 40, 43-56.

Messing, Joachim, *et al*. (1977). Filamentous coliphage M13 as a cloning vector vehicle: insertion of a *Hind*II fragment of the *lac* regulatory region in M13 replicative form in *vitro*. *Proceedings of the National Academy of Science of the United States of America*, 74(9), 3642-3646.

Nercessian, Olivier, *et al*. (2005). Bacterial populations active in metabolism of C_1 compounds in the sediment of Lake Washington, a freshwater lake. *Applied and Environmental Microbiology*, 71(11), 6885-6899.

Nübel, Ulrich, *et al*. (1999). Quantifying Microbial Diversity: Morphotypes, 16S rRNA Genes, and Carotenoids of Oxygenic Phototrophs in Microbial Mats. *Applied Environmental Microbiology*, 65(2), 422-430.

National Resource for Molecular Biology Information. (2007). *Nucleic Acid Research*, 35(Database issue):D21-5. http://www.ncbi.nih.gov/Genbank.

Olsen, Gary J., *et al*. (1986). Microbial Ecology and Evolution; A Ribosomal RNA Approach. *Annual Review of Microbiology*, 40, 337-365.

Olsen, Gary J., and Woese Carl R. (1993). Ribosomal RNA: A Key to Phylogeny. *The FASEB* Journal, 7, 113-123.

Pace, Norman P. (2008). The Molecular Tree of Life Changes How We See, Teach, Microbial Diversity. *Microbe*, 15-20.

Promega Corporation. (2007). *EcoR1 Enzyme Restriction Digest Usage Information*. Madison, WI.

QIAGEN. (2006). *QIAprep® Miniprep Handbook*. Second edition. Valencia, CA.

Retief, Jacques D. (1999). Bioinformatics Methods and Protocol by Stephen Misener and Stephen A. Krawetz. Chapter 12: Phylogenetic analysis using PHYLIP, 243-258. New Jersey: Humana Press.

Ribosomal Database Project Staff. (2007). Personal communications.

Schloss, P.D. & Handelsman, J. 2005. Introducing DOTUR, a Computer Program for Defining Operational Taxonomic Units and Estimating Species Richness. *Applied and Environmental Microbiology*, 71(3):1501-1506.

Schloss, Patrick & Handelsman, Jo (2006). Toward a Census of Bacteria in Soil. *PLoS Computational Biology*, 2(7), 1-13.

Servaites, Jerome. (2007). PhD, Wright State University Manager, EEE Genomics Laboratory, College of Science and Math, Wright State University, Dayton, OH.

Smith, Stephanie. (2007). PhD, Battelle. Personal Correspondence.

Sogin, Mitchell L., *et al*. (2006). Microbial diversity in the deep sea and the underexplored ‘‘rare biosphere’’. *Proceeding of the National Academy of Science of the United States of America*, 103(32), 12115–12120.

Stottmeister, U., *et al*. (2003). Effects of plants and microorganisms in constructed wetlands for wastewater treatment. *Biotechnology Advances*, 22, 93-117.

Stratagene®. (2007). *StrataClone™ PCR Cloning Kit Instruction Manual*. La Jolla, CA

Ward, C.H., Cherry J.A., & Scalf M.R. (1997) Subsurface Restoration. Chelsea, Michigan: Ann Arbor Press Inc.

Wang, Q, G. M. Garrity, J. M. Tiedje, and J. R. Cole. (2007). Naïve Bayesian Classifier for Rapid Assignment of rRNA Sequences into the New Bacterial Taxonomy. *Applied Environmental Microbiology*, 73(16), 5261-5267.

Watanabe, Kazuya, *et al*. (2001). Design and Evaluation of PCR Primers to Amplify Bacterial 16S Ribosomal DNA Fragments Used for Community Fingerprinting. *Journal of Microbiological Methods*, 44, 253-262.

Woese, Carl R. & Fox, George E. (1977). Phylogenetic structure of the prokaryotic domain: The primary kingdoms. *Proceedings of the National Academy of Science*, 74(11), 5088-5090.

Woese, Carl R. (1987). Bacterial Evolution. *Microbiological Reviews*, 51(2), 221-271.

Wong, Karen M., *et al*. (2008). Alignment Uncertainty and Genomic Analysis. *Science*, 319, 473-476.

Yan, Jun. (2006). "Evaluation of Chlorinated Solvent Removal Efficiency Among Three Wetland Plant Species: A Mesocosm Study." Dept of Systems Systems and Engineering Management (M.S. ed). Wright-Patterson AFB, OH: Air Force Institute of Technology.

Vita

Captain Elisabeth M. Leon graduated from Clarksville High School in Clarksville, Tennessee. She entered undergraduate studies at the University of Tennessee in Knoxville, Tennessee where she graduated with a Bachelor of Science degree in Civil Engineering in December 2003. She was commissioned through the Detachment 300 AFROTC at the University of Tennessee where she was nominated for a Regular Commission.

Her first assignment was at Maxwell AFB as a student in the Aerospace and Basic Course in January 2004. In March 2004, she was assigned to 27th Civil Engineer Squadron, Cannon AFB, New Mexico where she served in the Maintenance Engineering section. While stationed at Cannon, she was reassigned to the 27th Fighter Wing as Chief of Protocol. In August 2006, she entered the Graduate School of Engineering and Management, Air Force Institute of Technology. Upon graduation, she will be assigned to the Civil Engineer and Services School at Wright-Patterson AFB, OH.

REPORT DOCUMENTATION PAGE	*Form Approved* *OMB No. 074-0188*

1. REPORT DATE *(DD-MM-YYYY)*	**2. REPORT TYPE**	**3. DATES COVERED** *(From – To)*
03-07-08	Master's Thesis	August 2006 – March 2008

4. TITLE AND SUBTITLE Molecular Characterization of Wetland Soil Bacterial Communities in Constructed Mesocosms	**5a. CONTRACT NUMBER** **5b. GRANT NUMBER** **5c. PROGRAM ELEMENT NUMBER**
6. AUTHOR(S) Leon, Elisabeth M., Captain, USAF	**5d. PROJECT NUMBER** **5e. TASK NUMBER** **5f. WORK UNIT NUMBER**
7. PERFORMING ORGANIZATION NAMES(S) AND ADDRESS(S) Air Force Institute of Technology Graduate School of Engineering and Management (AFIT/EN) 2950 Hobson Way, Building 640 WPAFB OH 45433-8865	**8. PERFORMING ORGANIZATION REPORT NUMBER** AFIT/GES/ENV/08-M04
9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES) This space intentionally left blank	**10. SPONSOR/MONITOR'S ACRONYM(S)** **11. SPONSOR/MONITOR'S REPORT NUMBER(S)**

12. DISTRIBUTION/AVAILABILITY STATEMENT

APPROVED FOR PUBLIC RELEASE; DISTRIBUTION UNLIMITED.

13. SUPPLEMENTARY NOTES

14. ABSTRACT

16S rRNA gene analysis was performed on soil samples from a subsurface flow mesosocm experiment, mimicking a constructed wetland. Three typical wetland species of plant were used in nine of the columns and three columns were unplanted controls. Analyses showed that plant presence, plant species, and depth did not have a significant effect on microbial community composition. However, richness analyses illustrated that plant presence had a positive effect on richness, plant species analyses indicated that there was evidence that plant species affected richness in a significant manner, and lastly richness was negatively affected in the bottom depth for the control and *Carex comosa* communities.

15. SUBJECT TERMS

Wetland; Diversity; Microbial Communities

16. SECURITY CLASSIFICATION OF:			**17. LIMITATION OF ABSTRACT**	**18. NUMBER OF PAGES**	**19a. NAME OF RESPONSIBLE PERSON** Bleckmann, Charles A. Civ, ENV
a. REPORT	**b. ABSTRACT**	**c. THIS PAGE**			**19b. TELEPHONE NUMBER** *(Include area code)* (937) 255-6565, ext 4721 (Charles.bleckmann@afit.edu)
U	U	U	UU	139	