RHIZOSPHERE METAGENOMICS OF THREE BIOFUEL CROPS

By

Jiarong Guo

A DISSERTATION

Submitted
to Michigan State University
in partial fulfillment of the requirements
for the degree of

Microbiology and Molecular Genetics – Doctor of Philosophy

2016
ABSTRACT
RHIZOSPHERE METAGENOMICS OF THREE BIOFUEL CROPS

By
Jiarong Guo

Soil microbes form beneficial associations with crops in the rhizosphere and also play a major role in ecosystem functions, such as the N and C cycles. Thus large-scale cultivation of biofuel crops will have a significant impact on ecosystem functions at least regionally. In recent years, advances in high throughput sequencing technologies have enabled metagenomics, which in turn opens new ways to access the unknown majority in microbiology but poses great challenges for data analysis due to the large data size and short read length of sequence data sets.

We generated about 1 Tb of shotgun metagenomic data from rhizophere soil samples of three biofuel crops: corn, switchgrass, and Miscanthus. My central goal is to devise methods to extract meaning from this rhizosphere metagenomic data, with a focus on N cycle genes since N is the most limiting resource for sustainable for biofuel production. I initially provide a review of gene-targeted methods for analyzing shotgun metagenomics.

In the second chapter I develop a method that improves the speed with which rRNA genes fragments can be found and analyzed in large shotgun metagenome data sets, thereby avoiding primer bias and chimeras that are problematic with PCR-based methods. I present a pipeline, SSUsearch, to achieve faster identification of short subunit rRNA gene fragments plus provides unsupervised community analysis. The pipeline also includes classification and copy number correction, and the output can be used in traditional amplicon downstream analysis platforms. Shotgun derived rhizosphere data from this pipeline yielded higher diversity estimates than amplicon data but retained the grouping of samples in ordination analysis. Our analysis confirmed the known bias against Verrucomicrobia in a commonly used V6-V8 primer set as well as discovered likely biases against Actinobacteria and for Verrucomicrobia in a commonly used V4 primer set.

In the third chapter, I explore an alternative phylogenetic marker to the widely used SSU rRNA gene, which has several limitations including multiple copies in the same genomes and low resolu-
tion for differentiating strains. I demonstrate that rplB, a single copy protein coding gene, provides finer resolution more akin to species and subspecies level and also finer scale (OTU) diversity analysis. The method requires shotgun sequence since the gene is not conserved enough for recovery by primers. When used on the rhizosphere sequence data, it revealed more microbial diversity and better differentiated the communities among the three crops than the SSU rRNA gene analysis.

In the last chapter I address my central biological question on rhizosphere metagenomics: do they differ among the three crops and what does this information suggests about function? I compare the rhizosphere metagenomes for overall community structure (SSU rRNA gene), overall function (annotation from global assembly), and N cycle genes (using Xander, a targeted gene assembly tool). All three levels showed corn had a significantly different community from Miscanthus and switchgrass (except for ammonia-oxidizing Archaea), and that the two perennials showed a trend of separation. In terms of life history strategy, the corn rhizosphere was enriched in copiotrophs while the perennials were enriched in oligotrophs. This is further supported by higher abundance of genes in the “Carbohydrates” subsystem category and higher fungi/bacteria ratios. Additionally, the nitrogen fixing community of corn was dominated by nifH genes most closely affiliated to Rhizobium and Bradyrhizobium while the perennials had nifH sequences most related to Coraliomargarita, Novosphingobium and Azospirillum, indicating that the perennials independently selected beneficial members. Moreover, higher numbers of nitrogen fixation genes and lower number of nitrite reduction genes suggest better nitrogen sustainability of the perennials. These data indicate that perennial bioenergy crops have advantages over corn in higher microbial species richness and functional diversity as well as in selecting members with beneficial traits, consistent with N use efficiency.
ACKNOWLEDGEMENTS

First, I am very grateful to have Dr. James Tiedje and Dr. Titus Brown as my co-advisers. They not only have provided continuous support of my whole Ph.D study with their great patience and immense knowledge in their fields (Microbiology and Data Science, respectively) but also tailor their mentoring styles based on my knowledge background and personality. I am especially grateful for the freedom they give me for independent critical thinking and exploratory research, which, combined with the joy from solving challenging problems, are the most enjoyable things for me in scientific research.

I also would like to thank my committee members (Dr. James Cole, Dr. Chris Adami, Dr. Sheng Yang He, Dr. Yanni Sun for their insight comments, encouragement, and hard questions in my comprehensive exam that widen my research from various perspectives and help me grow as a scientist.

My sincere thanks also goes to my fellow lab mates in both Brown lab and Tiedje lab for stimulating discussions and friendships.

Last but not the least, I would like to thank my parents for supporting me spiritually throughout my Ph.D study and my life in general. The wisdom and capacity to sympathize they have taught me are key to overcoming adversities during this journey.
# TABLE OF CONTENTS

<table>
<thead>
<tr>
<th>LIST OF TABLES</th>
<th>ix</th>
</tr>
</thead>
<tbody>
<tr>
<td>LIST OF FIGURES</td>
<td>x</td>
</tr>
</tbody>
</table>

## CHAPTER 1 REVIEW OF GENE-TARGETED METHODS IN SHOTGUN METAGENOMICS

1.1 Introduction ........................................................................................................... 1
1.1.1 Shotgun metagenomics .................................................................................... 1
1.1.2 Assembly graphs .............................................................................................. 2
1.1.3 Global assembly ............................................................................................... 3
1.2 Gene-targeted methods in shotgun metagenomics ................................................. 3
1.2.1 SSU rRNA gene ............................................................................................... 4
  1.2.1.1 EMIRGE ..................................................................................................... 4
  1.2.1.2 REAGO ..................................................................................................... 5
  1.2.1.3 SSUsearch ................................................................................................. 6
1.2.2 Functional genes ............................................................................................... 7
  1.2.2.1 Xander ....................................................................................................... 7
  1.2.2.2 SAT-Assembler .......................................................................................... 8
1.3 Significant issues with gene-targeted methods ..................................................... 9
  1.3.1 Reference dependency ..................................................................................... 9
  1.3.1.1 Lacking reference sequences for functional genes ..................................... 10
  1.3.1.2 Large sequence variation among subgroups within certain genes ............... 10
1.3.2 Short read vs. contig ......................................................................................... 11
1.3.3 Difficulty of read-based method for some functional genes ............................... 12
1.4 Summary and outlook ........................................................................................... 12

## CHAPTER 2 MICROBIAL COMMUNITY ANALYSIS USING RIBOSOMAL FRAGMENTS IN SHOTGUN METAGENOMES

2.1 Abstract ................................................................................................................. 14
2.2 Introduction ........................................................................................................... 14
2.3 Materials and Methods ......................................................................................... 15
2.3.1 Soil samples, DNA extraction and sequencing .................................................. 16
2.3.2 Data preprocessing ......................................................................................... 16
2.3.3 Building SSU and LSU rRNA gene models ...................................................... 17
2.3.4 Identification of rRNA gene fragments from metagenomic data ...................... 17
2.3.5 Testing the effect of target region size and variable region on clustering .......... 18
2.3.6 Community structure comparison based on OTUs from de novo clustering ........ 19
2.3.7 Comparison of OTU-based microbial community structures inferred from shotgun and amplicon SSU rRNA gene sequences ........................................... 19
2.3.8 Comparison of taxonomy based microbial community structures inferred from shotgun and amplicon SSU rRNA gene sequences ........................................... 20
2.3.9 Copy number correction ................................................................................... 20
2.3.10 Implementation, reproducibility and sequence data accession .............................................. 21
2.4 Results ........................................................................................................................................... 21
2.5 Discussion ...................................................................................................................................... 37
2.6 Conclusion ....................................................................................................................................... 43

CHAPTER 3 GENOMIC AND METAGENOMIC COMPARISON OF \textit{rplB} AND SSU rRNA GENE IN SPECIES RESOLUTION ..................................................................................................................... 45
3.1 Introduction ...................................................................................................................................... 45
3.2 Methods .......................................................................................................................................... 46
3.2.1 Downloading bacterial genomes ................................................................................................ 46
3.2.2 SSU rRNA gene and \textit{rplB} extraction and pairwise comparison .............................................. 46
3.2.3 Rhizosphere soil metagenomic data download and preprocess .................................................. 47
3.2.4 SSU rRNA gene and \textit{rplB} identification and comparison in soil metagenomes ...................... 47
3.3 Results ............................................................................................................................................ 49
3.3.1 SSU rRNA gene and \textit{rplB} copy number and intra-genome SSU identity ............................... 49
3.3.2 Pairwise comparison of genomes with SSU rRNA gene and \textit{rplB} .......................................... 50
3.3.3 Comparison of SSU rRNA gene and \textit{rplB} on diversity analyses in large soil shotgun metagenomes ......................................................................................................................... 55
3.4 Discussion ...................................................................................................................................... 56
3.4.1 Choosing genomic datasets ........................................................................................................ 58
3.4.2 Advantages of \textit{rplB} over SSU rRNA gene as a phylogenetic marker ..................................... 59
3.4.3 Novelties in this study .................................................................................................................. 60
3.4.4 Lower strain level resolution of SSU rRNA gene due to its intra-genome variations .................. 60
3.4.5 Comparison of two genes in \textit{de novo} OTU based community diversity analyses ................ 61
3.4.6 Sequencing depth needed for \textit{rplB} in metagenomes ................................................................ 61
3.4.7 Lack of universal primers for amplification and reference sequences ....................................... 62
3.5 Conclusion ...................................................................................................................................... 62

CHAPTER 4 RHIZOSPHERE METAGENOMICS OF THREE BIOFUEL CROPS ............................................ 64
4.1 Abstract .......................................................................................................................................... 64
4.2 Introduction .................................................................................................................................... 65
4.3 Methods ........................................................................................................................................ 67
4.3.1 Soil samples, DNA extraction, and sequencing ........................................................................ 67
4.3.2 Data preprocessing .................................................................................................................... 67
4.3.3 Diversity analysis using SSU rRNA gene ................................................................................... 69
4.3.4 Global metagenome assembly, annotation and functional diversity ......................................... 69
4.3.5 \textit{N} cycle gene assembly and diversity ...................................................................................... 70
4.3.6 Data accession ............................................................................................................................ 70
4.4 Results .......................................................................................................................................... 71
4.4.1 SSU rRNA gene based analysis ................................................................................................. 71
4.4.2 Global assembly ........................................................................................................................ 76
4.4.3 Functional diversity using annotation from assembly ............................................................... 76
4.4.4 \textit{N} cycle gene diversity ........................................................................................................... 78
<table>
<thead>
<tr>
<th>Section</th>
<th>Title</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>4.4.5</td>
<td>N cycle gene abundance and taxonomy composition</td>
<td>78</td>
</tr>
<tr>
<td>4.4.6</td>
<td>Metadata</td>
<td>84</td>
</tr>
<tr>
<td>4.5</td>
<td>Discussion</td>
<td>84</td>
</tr>
<tr>
<td>4.6</td>
<td>Conclusion</td>
<td>93</td>
</tr>
</tbody>
</table>

BIBLIOGRAPHY | 94 |
LIST OF TABLES

Table 1.1 Summary of gene-targeted methods for shotgun metagenomes . . . . . . . . 13
Table 2.1 Comparison of search results from SSUsearch, metaxa and meta-rna . . . . . 24
Table 2.2 Higher fungi/bacteria ratio and percent of AMF fungi are in rhizosphere sam-
ple (M1) than bulk sample (SB1) . . . . . . . . . . . . . . . . . . . . . . . . . . . 37
Table 3.1 Rhizosphere soil shotgun metagenomic data . . . . . . . . . . . . . . . . . . 48
Table 3.2 Summary of SSU rRNA gene and rplB copy number in genomes . . . . . . . 50
Table 3.3 Summary of reads identified as SSU rRNA gene and rplB in soil metagenomes . 56
Table 4.1 Size and JGI project ID of raw read data for 21 samples . . . . . . . . . . . . . 71
Table 4.2 Summary of SSUsearch pipeline results for 21 samples . . . . . . . . . . . . . 72
Table 4.3 Shared OTUs, genera, and contig among rhizosphere microbial communities
of corn (C), Miscanthus (M), and switchgrass (S) . . . . . . . . . . . . . . . . . 73
Table 4.4 Average percentage of Bacteria, Archaea, Eukaryota, and fungi in rhizosphere
communities of corn (C), Miscanthus (M) and switchgrass (S) . . . . . . . . . 75
Table 4.5 Global assembly statistics . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 76
Table 4.6 Soil chemistry test results for 21 samples . . . . . . . . . . . . . . . . . . . . 88
# LIST OF FIGURES

<table>
<thead>
<tr>
<th>Figure</th>
<th>Description</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>2.1</td>
<td>Flowchart of SSUsearch pipeline</td>
<td>23</td>
</tr>
<tr>
<td>2.2</td>
<td>Testing the effect of target region size and variable region on clustering</td>
<td>25</td>
</tr>
<tr>
<td>2.3</td>
<td>Testing the effect of target region size on V4 of full-length SSU rRNA genes in clustering from a synthetic community with 64 species</td>
<td>26</td>
</tr>
<tr>
<td>2.4</td>
<td>Technical reproducibility test of our unsupervised clustering and comparison of OTU abundances between paired shotgun and amplicon data</td>
<td>27</td>
</tr>
<tr>
<td>2.5</td>
<td>Phyla of OTUs significantly different between shotgun data and amplicon data</td>
<td>29</td>
</tr>
<tr>
<td>2.6</td>
<td>Principle coordinates analysis (PCoA) of amplicon and shotgun derived data from seven field replicates</td>
<td>30</td>
</tr>
<tr>
<td>2.7</td>
<td>Comparison of ordination analysis with different variable regions</td>
<td>31</td>
</tr>
<tr>
<td>2.8</td>
<td>Alpha diversity comparisons between corn and Miscanthus using V2, V4, V6 and V8 regions</td>
<td>32</td>
</tr>
<tr>
<td>2.9</td>
<td>Bacterial phylum profile comparison using different variable regions</td>
<td>33</td>
</tr>
<tr>
<td>2.10</td>
<td>Taxonomy profiles of Bacterial phyla of shotgun fragments from SSU and LSU rRNA genes and of amplicon reads</td>
<td>34</td>
</tr>
<tr>
<td>2.11</td>
<td>Taxonomy profile comparison at domain level using SSU and LSU rRNA genes</td>
<td>35</td>
</tr>
<tr>
<td>2.12</td>
<td>Bacterial phylum level taxonomy summary before and after SSU rRNA gene copy correction</td>
<td>36</td>
</tr>
<tr>
<td>2.13</td>
<td>Ordination plot of amplicon and shotgun derived data after copy correction</td>
<td>38</td>
</tr>
<tr>
<td>2.14</td>
<td>Length distribution of trimmed reads after quality trimming and paired-end merging</td>
<td>40</td>
</tr>
<tr>
<td>3.1</td>
<td>Intra-genome variation of SSU rRNA gene of R. leguminosarum, P. putida, and E. coli in boxplot</td>
<td>49</td>
</tr>
<tr>
<td>3.2</td>
<td>SSU rRNA gene and rplB identities among species in the same genus</td>
<td>51</td>
</tr>
<tr>
<td>3.3</td>
<td>SSU rRNA gene and rplB identities among R. leguminosarum related genomes</td>
<td>52</td>
</tr>
<tr>
<td>3.4</td>
<td>SSU rRNA gene and rplB among P. putida related genomes</td>
<td>53</td>
</tr>
<tr>
<td>3.5</td>
<td>SSU rRNA gene and rplB identities among E. coli related genomes</td>
<td>54</td>
</tr>
</tbody>
</table>
Figure 3.6  SSU rRNA gene and rplB identities among all E. coli genomes ............... 55
Figure 3.7  Comparison of OTU numbers using SSU rRNA gene and rplB in large soil metagenomes ................................................................. 57
Figure 3.8  Comparison of SSU rRNA gene and rplB in beta diversity analysis (ordination) using large soil metagenomes ........................................... 57
Figure 3.9  Comparison of SSU rRNA gene and rplB in alpha diversity analysis (Shannon diversity index) using large soil metagenomes .................... 58
Figure 4.1  Roots (Miscanthus) and rhizosphere soil ........................................... 68
Figure 4.2  Species (OTU) and functional diversity comparision among three crops .... 73
Figure 4.3  Rank abundance curve of microbial communities of corn (C), Miscanthus (M) and switchgrass (S) using OTU from SSU rRNA gene ..................... 74
Figure 4.4  Relative abundance of top ten most abundant bacterial phyla in three crops ... 75
Figure 4.5  Functional diversity comparison among three crops using subsystem level 1 ... 77
Figure 4.6  Function comparison between corn and Miscanthus ............................. 79
Figure 4.7  Ordination analysis of each N cycle gene and rplB using OTU ................. 80
Figure 4.8  OTU number comparison among corn (C), Miscanthus (M) and switchgrass (S) using OTU tables of N cycle genes ................................. 81
Figure 4.9  Diversity (Shannon) comparison among corn (C), Miscanthus (M) and switchgrass (S) using OTU tables of N cycle genes ............................. 82
Figure 4.10  Denitrification gene abundances after combining the genes encoding enzymes with the same function ......................................................... 83
Figure 4.11  Ratios of interest related to N cycle genes in corn (C), Miscanthus (M) and switchgrass (S) ................................................................. 83
Figure 4.12  Abundance and phylum level association of the indicated N cycle genes in corn (C), Miscanthus (M) and switchgrass (S) ................................. 84
Figure 4.13  Abundance and genus level distribution of nifH ................................. 85
Figure 4.14  Sequence identity of assembled N cycle genes to their best-hit references ... 86
Figure 4.15  Distribution of identity of assembled nifH sequences to their best-hit references ......... 87
Figure 4.16  Ratio of 2012 yield over 2011 and N₂O flux measured near sampling time .... 87
Figure 4.17 Species (OTU) diversity of three crops
CHAPTER 1

REVIEW OF GENE-TARGETED METHODS IN SHOTGUN METAGENOMICS

1.1 Introduction

Sequencing technology has advanced from Sanger sequencing to massive short read ("next" or second generation) sequencing such as 454 and Illumina, and then again to massive long read (third generation) sequencing such as PacBio and Oxford Nanopore, just in the past ten years [1, 2]. Second generation sequencing technologies provide very high throughput and acceptable sequencing error, and has transformed biology research because of their low cost and large amounts of data. Third generation sequencing technologies are promising to provide long read alternatives to second-generation technology, at the cost of higher sequencing error rates [3].

1.1.1 Shotgun metagenomics

Metagenomics, the direct sequencing of DNA extracted from environmental samples, provides new ways to access the uncultured majority in microbiome studies. Techniques based on amplicon sequencing or gene-targeted metagenomics are limited by the lack of universal primer sets and the presence of both primer bias and PCR artifacts [4–6]. Since metagenome shotgun sequencing samples from all genomes in the community, it can potentially overcome these problems, providing a more complete picture of the microbial community. Massive short read sequencing of shotgun metagenomes has advanced our knowledge about microbial community members in many environments [7, 8].

With the large data sets produced from second generation data, data analysis become the main challenge in many research projects [9, 10]. There are two main methods for first-stage analysis of short read data: read mapping and assembly. Read mapping requires a reference sequence (genomes and genes) while assembly does not; assembly is a process that join reads of fragmented DNA to recover the sequence of the original full length DNA (usually chromosomes or genomes).
For several reasons, metagenome assembly generally does not produce complete genomes, but just longer sequences called contigs [7, 9], which could be further merged into scaffolds by paired-end or mate-pair information [11].

1.1.2 Assembly graphs

Assembly outputs are usually linear sequences but the assembly process requires a connectivity graph and algorithms that operate on that graph to produce assembled contigs. Two of the most common data structures used for the assembly process are de Bruijn graph (DBG) and string graph (SG). The de Bruijn graph first chops reads into even smaller k-mers and then builds a graph connecting k-mers that share k-1 bases, while the string graph first finds overlaps (larger than a length cutoff) among all reads and then connect reads based on the overlapping information [11, 12]. In the string graph, overlap calculation requires all-against-all pairwise read comparison and thus is very expensive in CPU time. The de Bruijn graph is unintuitive because it breaks down the reads but it achieves faster CPU time by avoiding the expensive all-against-all pairwise comparisons. There are only eight possible neighboring k-mers for each k-mer by extending a base (A, T, C, or G) on either side and neighboring k-mer connections can be calculated very efficiently just by checking whether those eight possible neighboring k-mers are present in the data. In terms of memory cost, the string graph increases with sequencing depth while the de Bruijn graph is not related to sequencing depth (only determined by total genome size) [13], if there are no sequencing errors. Both string graph and de Bruijn graph are sensitive to sequencing errors, which greatly increases the complexity of the graphs and memory usage. Each sequencing error can cause k spurious nodes (k-mers) in de Bruijn graph and one spurious node (string) in string graph.

The string graph works well with long reads by preserving the integrity of reads but does not scale with high depth data, whereas the de Bruijn graph fits well with high depth coverage short reads that second generation sequencing platform produces (454 and Illumina). This resulted in a boom of de Bruijn graph assemblers, such as Velvet, ABySS, AllPaths and SOAPdenovo
Alternatively, Simpson et al. [12] have applied FM index derived from Burrows-Wheeler transformation to find overlaps among reads which reduces the time complexity of string graph assembly to be linear with genome size and independent of the sequencing depth. This makes string graphs an option with second generation sequencing.

1.1.3 Global assembly

In analyzing shotgun metagenomes, an intuitive approach is to first recover all genomes in the metagenomes, i.e., global assembly in contrast to gene-targeted (local) assembly. Global assembly has become a common step for shotgun metagenomic analyses and there are many metagenome assemblers available such as MetaVelvet, IDBA-UD, and MEGAHIT [17–19]. Although there have been many major improvements made in recent years with read utilization in assemblies [17], global assembly results are still limited by challenges such as repeats, sequencing errors, low coverage and large data sets, especially for complex environments like soil [7]. Moreover, global assembly does poorly on genes with conserved regions such SSU rRNA genes due to the complexity this introduces in the assembly graph; typically, global assembly collapses sequences with high identities that are useful for diversity analysis [6, 20].

1.2 Gene-targeted methods in shotgun metagenomics

Many studies only focus on important genes or pathways such as the nitrogen cycle, biodegradation or antibiotic resistance genes so global assembly is not necessary for these studies. Gene-targeted methods can also be more computationally efficient because they focus only on reads identified as the target gene. Here we discuss several gene targeted methods that are assembly- or read-based and group them by their desired output – SSU rRNA gene or functional (protein-coding) gene.
1.2.1 SSU rRNA gene

SSU rRNA gene is the standard phylogenetic marker for diversity analyses [21, 22]. It also has the most comprehensive reference sequence databases (RDP and SILVA) which are important for taxonomy and diversity analyses [23, 24]. Thus it is a gene of interest for most metagenomic studies. There are several tools developed to extract the SSU rRNA gene fragments and/or assemble them into full length for further diversity analyses [6, 20, 25].

1.2.1.1 EMIRGE

SSU rRNA genes are difficult to assemble using global assembly methods because they have both highly conserved regions and also hyper variable regions; this in turn forms a complex graph structure that it is hard for assemblers to disentangle. EMIRGE [26] overcomes this problem by avoiding assembly, and uses a mapping-based strategy instead. It applies an expectation maximization (EM) algorithm to recover full length SSU rRNA gene sequences. It takes raw reads in FASTQ format (either paired ends or single end), a SSU rRNA gene reference database in FASTQ format, and a Bowtie index of the reference database as input. First, reads are mapped to the references by Bowtie [27]. Second, EMIRGE runs an EM algorithm. The expectation step calculates the probability a read is from a reference sequence and the maximization step modifies the reference sequences based on the most frequent base at each position from read mapping while also calculating the relative abundance of each reference sequence. Then the above mapping, expectation, and maximization steps are repeated until the reference sequences and their relative abundances stabilize. By default, EMIRGE stops after 40 iterations if the reference sequences and relative abundances do not stabilize. For complex communities, the authors recommend increasing the number of iterations. During an iteration, two references are merged if they become very similar and their identity exceeds a certain threshold (default is 0.97). On the other hand, a sequence is split into two if the variant fraction exceeds a threshold (default of 0.04) based on read mapping, suggesting two different strains have the best hit on the same reference.

The most important contribution of EMIRGE is that it provides a probabilistic framework
to address the uncertainty of mapping SSU rRNA gene fragments to the reference sequences. For example, some fragments from conserved regions could map to many references, and even fragments from variable regions could map to many references if multiple closely related strains are present. This probabilistic framing enables references to merge and to diverge rather than be kept as composite or consensus sequences from different strains as the global assemblers do. The final sequence output also provides a probability that can guide user confidence in the results.

As with global assembly tools, EMIRGE also sometimes produces chimeric sequences [28]. Erroneously chimeric sequences in the reference database might be passed on to the reconstructed SSU rRNA genes if there are reads mapped across the chimeric reference. Another scenario is that SSU rRNA gene sequences from closely related strains but with low coverage might form chimeras because there are not enough reads mapped to their closest reference; in this case, EMIRGE does not have enough information on variant fractions to split the reference into two or more candidate sequences.

1.2.1.2 REAGO

REAGO [25] is another method for reconstructing SSU rRNA content from a sample, but it is assembly based, unlike EMIRGE. REAGO first identifies the SSU rRNA gene reads with a profile called covariance model (CM) using INFERNAL, which is aware of secondary structure [29]. Then the identified SSU rRNA gene reads are used to build a string graph by finding overlaps among reads. This string graph is reduced with sequencing error correction, node-collapsing, topology-based pruning and taxonomy-based removal of bad edges. Sequences are then assembled by a path finding algorithm guided by paired-end information to avoid chimeras. Finally, partial assemblies shorter than the length cutoff are merged into longer ones by paired-end information.

REAGO uses string graphs rather than de Bruijn graphs since the multiple conserved regions and variable regions in the SSU rRNA gene cause the graphs to be very complex. Because De Bruijn graphs break reads down into k-mers, they deal poorly with this kind of complexity.

Although this string graph approach is less prone to chimeras than de Bruijn graphs, the SSU
rRNA gene is still a particularly difficult challenge for string graph assembly. This is because reads sharing overlaps are not necessarily from the same gene sequence due to the high conservation of the SSU gene. REAGO utilizes taxonomy information from each node along with paired-end information to find the correct paths for assemblies and reduce chimeras. Despite the advanced algorithm used to resolve chimeras, REAGO is very slow in the search CM search step.

1.2.1.3 SSUsearch

Since the SSU rRNA gene is difficult to assemble, there are also alternatives that avoid assembly and do diversity analyses directly on the short reads from shotgun sequence. Early pilot amplicon-based studies used 80 bp reads, which is sufficient for common microbial diversity analyses [30]. But, as Illumina reads increased in length past 300 bp, merging paired ends yields sequences that are comparable to amplicon sequences in their length and specificity.

SSUsearch [6] is a tool that uses shotgun reads directly for diversity analyses. It includes two main steps. The first step is fast identification of SSU rRNA gene fragments using HMMER, which shows significant speed and memory improvement compared to other tools such as BLAST, ribopicker, and Meta-rna [31–33]. The second step is the selection of reads mapped to a specific region (V4, E. coli position: 577 to 727 by default) so all reads have overlaps with each other. Then those reads can be further used for typical amplicon-like diversity analyses, e.g. as provided in Mothur and QIIME [34, 35].

The main advantage of SSUsearch is that it enables de novo OTU (Operational Taxonomic Unit) based diversity analyses with Illumina shotgun sequencing data. Secondly, SSUsearch is designed to scale with large datasets, e.g., 5 CPU hours and less than 5 Gb memory on search step (the bottle neck step) for 38 Gb (one lane) of trimmed Illumina HiSeq2500 data. SSUsearch makes a speed improvement by merging bacteria and archaea models since the merged model gives almost the same results as two separate models with their outputs combined. Second, there is a large memory improvement by code optimization, compared to Meta-rna [33], commonly used and HMM-based tool for SSU rRNA gene identification. The search step (HMMER) of SSUse-
arch, however, has not been compared to the one of REAGO (INFERNO). It is well known that INFERNO (cmsearch) is much slower and might have higher sensitivity and specificity on search results than HMMER (hmmsearch) but no tests have been done on simulated and real metagenomic data.

1.2.2 Functional genes

Since functional genes can be analyzed in protein space, analyses such as homology search and gene alignment are done with different tools than those for SSU rRNA genes (analyzed in nucleotide space). Moreover, protein domains are commonly not unique to one gene and often shared among many different genes, thus directly aligning short reads may not work well because reads from other genes sharing a protein domain can be incorrectly assigned. Therefore assembly-based methods are more suitable due to higher confidence provided by longer assembled sequences.

1.2.2.1 Xander

Xander [36] is a gene-targeted de Bruijn graph assembler that uses Hidden Markov Models (HMMs) of protein sequences of the targeted gene to guide graph traversal. Xander first loads reads into a de Bruijn graph using a low-memory probabilistic data structure called a Bloom filter, which is commonly used to reduce the memory cost of large sequence data [10]. Then it finds k-mers in nucleotide reference sequences of target genes in the graph as the starting point for graph traversal. A forward HMM and reverse HMM built respectively from the aligned protein sequences and the reverse of the aligned protein sequences enables graph traversal in both directions from starting k-mers. Guided by the protein HMM, the graph traversal advances three bases at a time, which equals to one codon. Xander use an A* algorithm to find the path with the highest score according the protein HMM. Further, the candidate assembled sequences are then clustered at 99 % to remove redundant sequences, and chimeras are removed by UCHIME [37]. Xander also has post-assembly utilities including assembled sequence coverage calculation and taxonomy assignment by nearest neighbor.
To run Xander, protein HMMs (both forward and reverse) and a large set of protein and nucleotide reference sequences with taxonomy information are required. Currently, Xander focuses on *rplB* (50S ribosomal subunit protein L2, a single copy house keeping gene) and N cycle genes including AOA, AOB, *nifH*, *nirK*, *nirS*, *norB_cNor*, *norB_qNor*, *nosZ_cladeI* and *nosZ_cladeII* are included in the tool. It is fairly easy to add more genes (https://github.com/rdpstaff/Xander_assembler#per-gene-preparation-requires-biological-insight).

In summary, Xander applies HMM search on top of the commonly used *de Bruijn* graph paradigm. HMM guided graph traversal not only reduced the search space compared to global assembly but also provides a probability measure on how likely a contig based on the model, and thus reduce chimeras. It also recovers from shotgun data more of the targeted genes and assembles them to longer lengths than do global assemblers [36].

### 1.2.2.2 SAT-Assembler.

Similar to Xander, SAT-Assembler also uses HMMs but it is a string graph based assembler. It is from the same group as REAGO and has a similar design that includes two main steps. The first step searches for target gene fragments in reads using HMM with HMMER3, which greatly reduces the input data size for the next step. The second step builds string graphs for each targeted gene and then assembles. The read alignment location information to the HMM model from the first step is used to help guide overlap calculation among reads. Multiple types of information such as paired end, coverage, and alignment location on the HMM are used to guide graph traversal and avoid chimeras. At the end, contigs are merged into scaffolds using paired end information.

To run SAT-Assembler, a file containing HMMs of targeted genes is required. The HMM for a specific gene can be built from aligned protein sequences of the gene by hmmbuild command in HMMER3. Additionally, SAT-Assembler also works with HMMs in the Pfam database, which has 16306 HMMs in version 30.0 and covers about 80% of proteins sequences in UniProtKB [38].

Both Xander and SAT-Assembler uses HMMs but in different ways. Xander loads all reads into the *de Bruijn* graph and then uses HMMs to guide the graphs traversal. Thus the graph traversal
is limited to a smaller part of the whole graph related to the targeted gene. The graph traversal space is reduced but it is still computationally expensive (CPU time and memory) to load all reads into the graph and identify the starting k-mers. On the other hand, SAT-Assembler only uses HMMs to identify the reads as target gene fragments, which are then used to build a much smaller graph. SAT-Assembler applies paired end and coverage information but not HMMs to guide graph traversal.

These two tools could be combined to make some improvements. The HMM search step provides the alignment location information to speed up the overlap calculation of string graph and also assists in establishing the correct connection among reads with low coverage. Although the search step is critical to reducing the problem size, it is still the bottleneck step of the SAT-Assembler. HMM based profile search provides a faster and more effective way to search the gene fragments compared to other types of search, e.g. pairwise alignment such as BLAST, because each HMM query is an aggregate profile formed from many sequences and thus the query size is greatly reduced. Additionally, HMM based profile search can improve the sensitivity of remotely related protein identification [39].

### 1.3 Significant issues with gene-targeted methods

#### 1.3.1 Reference dependency

For all the tools mentioned above, a certain type of reference information is required, either HMM, CM, or a collection of gene sequences with associated taxonomy information. Thus the quality and number of the references is key to the performance of these tools. For the SSU rRNA gene, millions of reference sequences have been collected in databases such RDP and SILVA [23, 24]; this provides rich phylogenetic context for diversity analyses such as taxonomy assignment and chimera check. In comparison, functional genes are still lacking in reference sequences.
1.3.1.1 Lacking reference sequences for functional genes

There are typically two categories of functional genes of interest, genes coding for critical ecosystem functions such as the N cycle, and universal protein coding genes that can serve as phylogenetic markers such as \textit{rplB}, \textit{rpoB} and \textit{recA}. The latter are of interest since they have higher resolution for species identification than SSU rRNA gene and are also single copy [40,41], unlike SSU rRNA genes. As microbial ecology moves beyond taxonomy and community structure to ecosystem functions, a large number of reference sequences for functional genes are also needed. Unlike SSU rRNA genes, it is more difficult to maintain databases of functional genes because there are too many of them and thus the effort has to be distributed to groups with special interest in certain genes. One effort in this direction is the RDP FunGene pipeline, which provides a platform for researchers to maintain functional gene databases in a semi-automatic way. As with the SSU rRNA gene, there are two main sources of reference sequences. The first source is whole length genes from genomes (about 60,000 in NCBI RefSeq when accessed in Sep 2016) and the second source is partial gene sequences amplified by primers (the majority). The number of sequenced genomes are accumulating fast since second generation sequencing started, so we are making good progress on the first source, although more efforts to extract genes for ecosystem functions in phylogenetically undersampled regions of the taxonomy are needed. For the second source, there are many previous studies targeting functional genes using primers [42–45]. A larger effort on database curation is needed to collect the references from the above two sources. In addition, new primer design is needed to find optimal primers for gene-targeted amplicon sequencing since this approach can sample the microbiome to greater depth. One successful example of such an effort is \textit{nifH}-coding for nitrogenase reductase [46,47].

1.3.1.2 Large sequence variation among subgroups within certain genes

Although functional genes share some conserved regions (domains), there is likely large variations among sequences, which could be further divided into subgroups (clades). Note that the SSU rRNA gene could also be divided into subgroups (phyla) but it is not usually necessary due to the high
sequence identity among all sequences. However, functional genes are much less conserved than the SSU rRNA gene. Thus alignment and homology search are more accurate when subgroups are treated separately.

There are two cases where gene targeted analysis is suboptimal. The first case is that we are missing representative sequences from some subgroups, which leads to biased results on diversity. The nitrous oxide reductase (\textit{nosZ}) is such an example: studies now show that there are at least two clades of \textit{nosZ} genes, with the second clade very understudied [48]. Follow-up studies show that clade II is very abundant in the environment and its enzyme is more efficient in low nitrogen environments [49]. The second case is that we are treating subgroups with significant sequence divergence as a group together, which can cause poor performances on gene search and sequence alignment. For example, the \textit{nosZ} clade I and clade II should be treated separately due to low sequence similarity between two groups and the combined sequences from two clades could lead to poor multiple sequence alignment and further a single \textit{nosZ} HMM built from the alignment with poor sensitivity and specificity (conserved domains are critical for HMM based search).

1.3.2 Short read vs. contig

Both short reads and assembled contigs are used in metagenomic studies [7, 9, 50], but the latter are more common, since annotating short reads is not scalable computationally and provides low confidence (low e-value or score in BLAST search). Additionally, assembled contigs can provide more information such as gene synteny and genomic information from novel species that can further be used for isolating those novel species. On the other hand, the assembled contigs still have some key drawbacks. First, there are chimeras in assembled contigs. Even though some can be resolved by paired end information, there is no method to verify them \textit{in silico} generally. Second, sequence variations from closely related strains are collapsed or ignored in the assembly process. Thus the assembled contigs are not suitable for SNP, primer design, or diversity analysis that involves fine taxonomy (subspecies, ecotype or strain) levels. Third, low coverage strains are often too low-coverage to assemble. All the above becomes a barrier in complex metagenomes from
environments such as soil that contains not only high diversity but many closely related strains, and many strains with low coverage. Therefore, contigs may not reliably represent the population, although the annotations from them do have high confidence (lower e-value for homology search). In contrast, read-based methods do not encounter any of the above three issues and thus are still valuable for shotgun metagenomics. Despite of the low confidence in annotation caused by short length, reads from second generation sequencing are shown to be long enough to give reliable annotations using HMM based profile search [51]. Additionally, the scalability of read-based annotation is not as a big issue in gene-targeted analysis as in global functional diversity analysis since only reads belonging to targeted genes are annotated.

1.3.3 Difficulty of read-based method for some functional genes

As mentioned above, conserved but not unique domains, such as sulfur and iron binding domains in \textit{nifH}, confound application of read based methods in functional genes, since reads from a different gene with the same domain can falsely increase the coverage and sequence variation. However, read-based methods are favored by avoiding those non-unique domains and looking at other regions in the target gene including gene specific domains and variable regions. Yet the challenges are that not all genes have unique domains and phylogenetically diverse reference sequences are required to search variable regions of a gene using BLAST style pairwise comparisons, which is not feasible for most functional genes currently.

1.4 Summary and outlook

In summary, we review a few tools for gene-targeted analyses with shotgun sequence (metagenomes) (Table 1.1). We also discuss some significant issues related to gene-targeted analyses with shotgun metagenomes, including the lack of reference sequence databases and the presence of unidentified subgroups for functional genes, and difficulty of applying short read based methods on functional genes due to widespread and highly conserved but non-specific domains.
Table 1.1: **Summary of gene-targeted methods for shotgun metagenomes.**

<table>
<thead>
<tr>
<th>Tool</th>
<th>Gene</th>
<th>Search</th>
<th>Assembly</th>
<th>Read/Contig</th>
<th>Diversity analysis</th>
</tr>
</thead>
<tbody>
<tr>
<td>EMIRGE</td>
<td>SSU</td>
<td>Bowtie</td>
<td>EM</td>
<td>Contig</td>
<td>No</td>
</tr>
<tr>
<td>REAGO</td>
<td>SSU</td>
<td>CM</td>
<td>SG</td>
<td>Contig</td>
<td>No</td>
</tr>
<tr>
<td>SSUsearch</td>
<td>SSU</td>
<td>HMM</td>
<td>HMM</td>
<td>Read</td>
<td>Yes</td>
</tr>
<tr>
<td>Xander</td>
<td>Protein</td>
<td>Kmer</td>
<td>DBG&amp;HMM</td>
<td>Contig</td>
<td>Yes</td>
</tr>
<tr>
<td>SAT-Assembler</td>
<td>Protein</td>
<td>HMM</td>
<td>SG</td>
<td>Contig</td>
<td>No</td>
</tr>
</tbody>
</table>

Here we mainly discuss *in silico* gene targeted analyses in shotgun data but it is also possible to target specific genes *in vitro* using probes designed to that gene and then sequence, which can give much higher coverage for low abundance genes [52]. The probe design (similar to primer design) will require a phylogenetically diverse set of reference sequences, or else the method will be biased. In the future, with the further development of long-read sequencing technologies (with reduced error rates), read-based methods will benefit from longer reads. Especially for a single gene, it might be not necessary to assemble with long reads. Another way to think about the problem is that long-read sequencing and assembly serves the same purpose, i.e., to obtain long sequences. Additionally, string graph may gain more popularity as occurred in Sanger sequencing era compared to the dominance of *de Bruijn* graph with short read data from second generation sequencing, which applies to both global assembly and gene-targeted assembly. Meanwhile, short-read sequencing and *de Bruijn* graph based methods will continue to play an important role in metagenomic analysis due to its high sequencing depth.
CHAPTER 2

MICROBIAL COMMUNITY ANALYSIS USING RIBOSOMAL FRAGMENTS IN
SHOTGUN METAGENOMES

2.1 Abstract

Shotgun metagenomic sequencing does not depend on gene-targeted primers or PCR amplification and thus is not affected by primer bias or chimeras. However, searching rRNA genes from large shotgun Illumina dataset is computationally expensive and there is no existing approach for unsupervised community analysis of SSU rRNA gene fragments retrieved from shotgun data. We present a pipeline, SSUsearch, to achieve faster identification of short subunit rRNA gene fragments and enabled unsupervised community analysis with shotgun data. It also includes classification and copy number correction, and the output can be used by traditional amplicon analysis platforms. Shotgun metagenome data using this pipeline yielded higher diversity estimates than amplicon data but retained the grouping of samples in ordination analyses. We applied to this pipeline to soil samples with paired shotgun and amplicon data, and confirmed bias against Verrucomicrobia in a commonly used V6-V8 primer set as well as discovering likely bias against Actinobacteria and for Verrucomicrobia in a commonly used V4 primer set. This pipeline can utilize all variable regions in SSU rRNA and can also be applied to large subunit rRNA (LSU) genes for confirmation of community structure. The pipeline can scale to large soil metagenomic data (5 Gb memory and 5 CPU hours to process 38GB (1 lane) of trimmed Illumina HiSeq2500 data) and is freely available at https://github.com/dib-lab/SSUsearch under the BSD License.

2.2 Introduction

Microbial phylogeny, identification and evolution studies were revolutionized by the introduction of SSU rRNA analysis 25 years ago [21], and with the advent of PCR and high throughput sequencing, community structure studies are now commonplace [22, 30, 53, 54]. The growing size of SSU rRNA gene databases provide a rich ecological and phylogenetic context for SSU rRNA gene-based community structure surveys [24, 55]. However, the accuracy of PCR-based amplicon approaches are reduced by primer bias and chimeras [5, 56].

Unlike gene-targeted amplicon sequencing, shotgun sequencing samples from the entire community, by sequencing randomly sheared fragments of DNA [9, 57]. Hence, while amplicon sequencing can provide far deeper coverage of SSU rRNA genes with same amount of sequencing, shotgun sequencing may provide a more accurate characterization of microbial diversity including functional diversity [58]. In particular, shotgun sequencing may provide an improved means to detect divergent sequences not recovered by standard SSU rRNA gene primers, such as the Verrucomicrobia, as well as eukaryotic members of the community [4, 56, 58, 59]. Note that both approaches remain prone to sequencing error and bias from environmental DNA extraction [5].

The challenges for using shotgun DNA for rRNA analyses are in efficiently searching for these fragments in large sequence data sets, and the subsequent analysis of the matching short reads. Several methods have been developed for SSU rRNA retrieval in large data sets [32, 33, 60, 61], but speed improvements are still needed to match the growth in data size; moreover, none of them provide further community analysis using the identified rRNA gene sequences. In addition, traditional community analysis tools [23, 34, 35] are largely designed to handle sequences that are amplified by PCR primers. There are two primary types of community analyses: reference-based (supervised) and OTU-based (unsupervised). The reference-based method assigns SSU rRNA gene sequences to bins based on taxonomy of their closest reference sequences, while OTU-based methods assign overlapping gene sequences to bins based on de novo clustering with a specified similarity cutoff (e.g. 97%). The reference-based method can be easily applied to shotgun data once SSU rRNA gene fragments are retrieved [62] and there are several existing tools available
[63–67], but the OTU-based approach still remains challenging with shotgun data because reads are from randomly sheared fragments.

The main goal of this study is to enable unsupervised OTU-based analysis of large shotgun metagenomic data from soil. We improved speed and memory efficiency with a Hidden Markov Model (HMM) based method, already shown to be fast and accurate for SSU rRNA search [33, 60, 61], using a well-curated and up-to-date training reference sequence collection from SILVA [24]. Our unsupervised clustering method was first tested on a synthetic community with shotgun data of 100 bp reads. We next applied the method to soil datasets, where we assembled longer reads from the overlapping paired end Illumina HiSeq reads and mapped those to 150 bp small hyper-variable regions of SSU rRNA genes for de novo clustering and further diversity analysis. We retrieved and analyzed the large ribosomal subunit (LSU) gene for confirmatory analysis. Finally, we evaluate primer biases using the paired shotgun and amplicon data produced from the same DNA extract, which is beyond traditional primer evaluation using in silico database search [68, 69].

2.3 Materials and Methods

2.3.1 Soil samples, DNA extraction and sequencing

Two sets of soil samples were used. The first sample, which was used to develop the method, was a bulk (non-root influenced) soil sample (SB1) taken in 2009 from between the rows of switchgrass. The method was then applied to the second sample set taken in 2012, which consisted of seven replicate rhizosphere samples from both corn (C) and Miscanthus (M) plots. All samples were from the Great Lakes Bioenergy Research Center (GLBRC) Cropping System Comparison site at the Kellogg Biological Station in Southwest Michigan, http://data.sustainability.glbrc.org/pages/1.html. The rhizosphere samples were closely associated with the roots, <1 mm. DNA extraction and SSU rRNA gene amplification methods were previously described [70]. The SSU rRNA gene amplicons from the first sample were sequenced by the Joint Genome Institute (JGI) in their standard work flow, which used 454 GS FLX and Titanium platforms and primer set
(926F: AAACTYAAAKGAATTGACGG 1392R: ACGGGCGGTGTGTRC) that targeted the V6-V8 variable region of Bacteria, Archaea, and Eukaryotes. The second set was also sequenced at JGI but at a later time, so used the Illumina MiSeq platform and primer set (515F: GTGCCAGCMGCGCGGTAA 806R: GGACTACHVGGGTWTCTAAT) that targeted the V4 variable region. Shotgun sequencing was also done by JGI using Illumina GAII platforms for the first set and HiSeq 2500 using 250 bp insert libraries and 2x150 bp reads for the second set. We had about 8 Gb of data for the first set and about 300 GB of data each for corn and Miscanthus for the second set.

2.3.2 Data preprocessing

Data preprocessing is necessary for both shotgun and amplicon data due to sequencing errors. However, it is not included as part of the core pipeline because users have their own preferences. We trimmed trailing bases with quality score 2 called Read Segment Quality Control Indicator encoded by ASCII 66 “B” in Illumina GAII or ASCII 35 “#” in Illumina HiSeq shotgun data and discarded reads shorter than 30 bp and with Ns. The reads were then quality trimmed with fastq-mcf (version 1.04.662) (http://code.google.com/p/ea-utils) with “-l 50 -q 30 -w 4 -x 10 –max-ns 0 -X”. The paired-end reads overlapping by more than 10bp were assembled into one long read by FLASH (version 1.2.7) [71] with “-m 10 -M 120 -x 0.20 -r 140 -f 250 -s 25”. Roche 454 pyrotag amplicon data was processed using the RDP Pipeline “PIPELINE INITIAL PROCESS” and “CHIMERA CHECK” [23]. Reads were sequenced from the reverse primer end to the forward primer end. Since the targeted region is about 467 bp (926F/1392R), most reads were not long enough to reach the forward primer. Thus only the reverse primer product was used for quality trimming. The minimum length was set to 400 bp and defaults were used for other parameters.

2.3.3 Building SSU and LSU rRNA gene models

We quality trimmed SILVA [24] SSU and LSU Ref NR database (version 115) sequences by discarding all sequences with ambiguous DNA bases and converting U to T. We then clustered them
at a 97% similarity cutoff using pick_otus.py (default with UCLUST) and pick_rep_set.py from QIIME (version 1.8.0) [35]. We chose the longest representative in each OTU to be further clustered at 80% similarity cutoff. We collected the longest sequence in each OTU, resulting in 4027 representative sequences for the SSU rRNA gene and 1295 for the LSU gene to obtain a phylogenetically diverse set of reference genes. We further grouped these sequences into two groups, one combining Bacteria and Archaea and the other containing only Eukaryota (also see Discussion). Each group was used to make two HMMs (Hidden Markov Model): one with sequences from previous step and the other with reverse complement, using hmmbuild in HMMER version 3.1 [39]. At last, the HMM files were concatenated into a single file for each gene. This step is not part of the pipeline and the resulting HMMs were included in the database of this pipeline.

2.3.4 Identification of rRNA gene fragments from metagenomic data

The analysis framework is shown in (Fig. 2.1), as well as the reasons for our choices of pipeline components. We searched Illumina shotgun metagenomic data with hmmsearch in HMMER version 3.1 [39] using the LSU and SSU HMM models. For testing the sensitivity of newly built models, we compared our tool with meta-rna [33] and metaxa [61]. We used an e-value of 10 for hmmsearch with the newly built HMMs. Meta-rna (rna_hmm3.py) (package was not versioned and latest version update on Oct. 21, 2011 was used) was run with flags “-k euk,bac,arc -e 0.00001”, and metaxa (metaxa_x) version 2.0.2 was run with flags “--allow_single_domain 1e-5,0 -N 1 -E 1e-5”. A bulk soil (SB1), a rhizosphere soil (M1) and a synthetic community sample [58] were used as test data. We aligned the HMMER hits from the e-value cutoff of 10 using the multiple sequence aligner “align.seqs” in Mothur (version 1.33.3) [72]. For SSU rRNA gene fragments, 18491 full-length SSU rRNA gene sequences (14956 from Bacteria, 2297 from Archaea, and 1238 from Eukaryota) from the SILVA database (release 102) [24] downloaded from the Mothur website (http://www.mothur.org/wiki/Silva_reference_files) were used as the template with flags “threshold = 0.5”, and “flip = t”. For LSU rRNA, Multiple Sequence Alignment (MSA) of representative sequences of SILVA LSU Ref NR database clustered at 97% similarity cutoff were
used as template with the same flags as for SSU. Based on alignment information provided in the “align.seqs” output report file, those shotgun reads with more than 50% mapped to a reference gene were designated as SSU rRNA or LSU rRNA gene fragments.

2.3.5 Testing the effect of target region size and variable region on clustering

We used shotgun data of a synthetic community comprised of 64 species, which were sequenced by paired end 100 bp method on Illumina HiSeq 2000 [58]. When testing target region size effect on clustering, we picked V4 with starting position at 577 in E.coli. Sizes from 50 to 180 bp with a 10 bp increment were chosen. The minimum read length was set to the target region size minus 5 bp if the region size was less than 100 bp, and 95 bp when the region size was longer than 100 bp. We used pre.cluster command in Mothur with 1 edit distance to collapse reads with errors and their original reads. Then de novo clustering was achieved by RDP McClust with upgma algorithm and minimum read overlapping length of 25 bp [23]. We chose McClust as the clustering tool due to its speed and memory efficiency [23, 73]. Next, we choose V2, V3, V4, V5, V6, and V8, starting at position 127, 427, 577, 787, 987, and 1227 in E.coli respectively [74], to test the hyper variable region effect on clustering result. Target region sizes of 80 bp and 120 bp and a distance cutoff of 5% were chosen. Further, the above analyses were also applied to 16S rRNA genes from the 64 species comprising the community to get the true OTU numbers.

2.3.6 Community structure comparison based on OTUs from de novo clustering

For the clustering analysis of shotgun and amplicon data, 150 bases corresponding to the V8 region (E. coli positions: 1227-1377) were aligned. Reads shorter than 100 bp were removed from the alignment and the remainder clustered using McClust with minimum overlap of 25 bp and upgma method (6). The clustering result was converted to the Mothur format and community structure comparison was done using the make.shared with label=0.05, dist.shared(calc=thetayc), and using the pcoa command. When comparing different regions, E. coli positions: 127-277, 577-727, and 997-1147 were chosen for V2, V4, and V6, respectively [74]. Procrustes analysis as implemented
in QIIME was used to transform V2, V6 and V8 PCoA results and minimize the distances between corresponding points in V4. The bulk soil sample (SB1) was sequenced in six lanes from one Illumina plate using DNA from the same extraction. We used these as technical replicates for testing the reproducibility of de novo OTU-based analysis on shotgun data. Since sequencing depth is critical for reproducibility testing, we pooled these into two samples of three lanes each, the first three lanes as SB1_123 and the remaining three lanes as SB1_456.

2.3.7 Comparison of OTU-based microbial community structures inferred from shotgun and amplicon SSU rRNA gene sequences

The abundance of each OTU in shotgun data and amplicon data (V6-V8 for SB1 and V4 for M1) from the same DNA extraction were compared to check the consistency between the two sequencing approaches. Pearson’s correlation coefficient and $R^2$ of linear regressions were used to evaluate the correlation between the two types of data and between technical replicates. All the abundances of each OTU were increased by a pseudo-count of one to allow display on a log scale (avoiding zeros).

2.3.8 Comparison of taxonomy based microbial community structures inferred from shotgun and amplicon SSU rRNA gene sequences

The SSU rRNA fragments from shotgun data and amplicon data were classified using RDP Classifier [62]. The reference SSU rRNA genes from RDP and SILVA are provided on the Mothur website and were used as training sets, with a bootstrap confidence cutoff for classification of 50%. Representative sequences of SILVA LSU Ref NR clustered at 97% similarity cutoff were used as training set with taxonomy information built from the sequence file for the LSU rRNA gene. The bacterial taxonomy profiles from shotgun data and amplicon data were compared at phylum level.
2.3.9 Copy number correction

We used SSU rRNA gene copy number database in Copyrighter [75] that provides copy number for each taxon in Greengenes database [76]. In the taxonomic summary, the abundance of each taxon is weighted by the inverse of its SSU rRNA gene copy number. Similarly in OTU-based analysis, the abundance of each OTU is weighted by the inverse of SSU rRNA gene copy number of its consensus taxon. Unclassified sequences are weighted by the inverse of average copy number of all taxa in the dataset.

2.3.10 Implementation, reproducibility and sequence data accession

The pipeline can be found in https://github.com/dib-lab/SSUsearch as a tutorial with ipython notebooks [77]. Scripts for reproducing the figures in this paper can be found at GitHub (https://github.com/dib-lab/2014-ssu-search/blob/master/analysis-in-paper.Makefile). The synthetic community data for testing can be downloaded from NCBI SRA under SRR606249. The amplicon data for C1-7 and M1-7 are deposited in JGI genome portal under project ID 1025756 with library ID M2094 and M2113 respectively, and SB1 deposited in NCBI under SRX902929. The shotgun data for the same three datasets are deposited in JGI portal (C1-7 are under project ID 1023764, 1023767, 1023770, 1023773, 1023776, 1023779, and 1023782; M1-7 are under project ID 1023785, 1023788, 1023791, 1023794, 1023797, 1023800 and 1018623; SB1 are under project ID 402775).

2.4 Results

We developed an optimized pipeline that readily analyses large data sets (Fig. 2.1). The pipeline has two major steps: SSU rRNA gene fragment search and unsupervised OTU analysis. HMMER-based methods search with HMMs and thus scales with increasing size of SSU rRNA gene database [23], so we chose them for the first search step. We used meta-rna [33] but could not run it on large data sets due to its poor memory management. We therefore simplified and optimized the approach
used by meta-rna. Since the search step is still the computational bottleneck (Fig. 2.1), our interest here was to make an improvement on search speed and memory efficiency while retaining accuracy. Our implementation is about 4 times faster and 100 times more memory efficient than meta-rna, and 10 times faster and 15 times more memory efficient than metaxa [61] (Table 2.1). The speed improvement is realized from two modifications: 1) We reduce the number of HMMs models to search with by merging Bacteria and Archaea models. SSU rRNA genes are highly conserved and thus the merged model still has high sensitivity. Even more so, we can increase the sensitivity by using more relaxed E-value cutoff, since false positives are tolerable in this initial search step. 2) We use reverse-complement HMMs rather than reverse complementing the reads, because the latter scales poorly with large datasets. The newly built HMM models (HMMs for Bacteria and Archaea together and HMMs for Eukaryota) cover most hits found by separate HMMs of the three domains (Bacteria, Archaea, and Eukaryota) in meta-rna in all three test data sets, the two soil data sets in this study and one synthetic community (Table 2.1). Our method identified 15566 (0.03%) of 44,787,632 quality trimmed and paired end merged sequences as SSU rRNA gene fragments in the bulk soil sample (SB1) and 112402 (0.04%) of 274,060,925 reads in the first Miscanthus replicate (M1).

Unsupervised OTU analysis with shotgun data is not available in any current pipeline and we develop a method for OTU clustering around a small region where all reads overlap. To show the validity of our unsupervised method, we did tests on effects of target region sizes and different variable regions with shotgun data from a synthetic community. We found all region sizes from 50 bp to 160 bp in V4 had an OTU number that approached the species number at a distance cutoff of 4% or 5% (Fig. 2.2) when testing target region size effect on OTU number. We also did a similar test with only full length SSU rRNA genes from 64 species to make sure OTU number is close to species number and confirmed that OTU number was close to species number (at a range from 50 to 60) when a cutoff of 0 to 0.06 was chosen (Fig. 2.3). When target region size was larger than 170 bp, the clustering tool (McClust) [23] did not cluster because percentage of non-overlapping reads exceeded its threshold. Based on the above results, we chose 80 bp or 120 bp as the target region
Figure 2.1: **Flowchart of SSUsearch pipeline.** SSU rRNA gene fragments were retrieved by an hmmsearch and alignment step, which could be further used for reference-based (supervised) diversity analysis (taxonomy). Those fragments aligned to 150 bp of a variable region could be used for OTU-based (unsupervised) diversity analysis. SSU rRNA gene identification hmmsearch is the most time consuming steps. For 1 lane of trimmed HiSeq data (38 Gb) from *Miscanthus* rhizosphere sample (M1), SSU rRNA gene identification took about 4 hours with peak memory usage of about 4.5 Gb. In the analysis pipeline, where there was not a clear performance difference between tools, we mostly used Mothur including databases. For (de novo) OTU based analysis, SSU rRNA gene reference sequences with taxonomy information are required for classification. Another smaller set of aligned references is required to align the gene fragments from shotgun data. The SILVA database (the official one, not the one included with QIIME or Mothur) was used to build the HMM since the reference set from SILVA was more up to date. Two scripts in QIIME were used to cluster (UCLUST, the default) and pick representative sequences. Building the HMM is not part of the pipeline but using the built HMM is. We found that complete-linkage clustering is faster and requires less memory with McClust than with Mothur (dist.seqs and cluster). Additionally, we use two scripts in QIIME (pick_otus.py and pick_rep_set.py) to select representative sequences for building the HMM due to ease of use and the GreenGenes database is included for use with the Copyrighter copy number correction tool.
Table 2.1: **Comparison of search results from SSUsearch, metaxa and meta-rna.** Subsets of 5 million reads from metagenome of a synthetic community of 48 Bacteria and 16 Archaea, SB1 (bulk soil metagenome) and M1 (*Miscanthus* rhizosphere metagenome) are used as testing data. Symbol “∩” indicates overlap.

<table>
<thead>
<tr>
<th></th>
<th>Mock</th>
<th>SB1</th>
<th>M1</th>
</tr>
</thead>
<tbody>
<tr>
<td>SSUsearch</td>
<td>6432</td>
<td>2789</td>
<td>2612</td>
</tr>
<tr>
<td>Metarna</td>
<td>6455</td>
<td>2781</td>
<td>2600</td>
</tr>
<tr>
<td>Metaxa</td>
<td>5322</td>
<td>2649</td>
<td>2444</td>
</tr>
<tr>
<td>SSUsearch∩metarna</td>
<td>6300</td>
<td>2759</td>
<td>2576</td>
</tr>
<tr>
<td>SSUsearch∩metaxa</td>
<td>5286</td>
<td>2642</td>
<td>2442</td>
</tr>
<tr>
<td>metarna∩metaxa</td>
<td>5304</td>
<td>2649</td>
<td>2436</td>
</tr>
<tr>
<td>SSUsearch∩metarna∩metaxa</td>
<td>5268</td>
<td>2642</td>
<td>2435</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>SSUsearch cpu time (min)</th>
<th>Meta-rna cpu time (min)</th>
<th>Metaxa cpu time (min)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>1.6</td>
<td>4.6</td>
<td>4.8</td>
</tr>
<tr>
<td></td>
<td>8.6</td>
<td>16.5</td>
<td>16.9</td>
</tr>
<tr>
<td></td>
<td>17.5</td>
<td>47.2</td>
<td>34.8</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>SSUsearch memory (Mb)</th>
<th>Meta-rna memory (Mb)</th>
<th>Metaxa memory (Mb)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>35</td>
<td>33</td>
<td>30</td>
</tr>
<tr>
<td></td>
<td>3406</td>
<td>4005</td>
<td>4234</td>
</tr>
<tr>
<td></td>
<td>452</td>
<td>456</td>
<td>572</td>
</tr>
</tbody>
</table>

size and 5% distance cutoff for testing different hyper variable regions. The number of OTUs created in all variable regions is close to the real number of species in the synthetic community except when using V3.
Figure 2.2: Testing the effect of target region size and variable region on clustering.
Figure 2.2 (cont’d): A shotgun metagenomic dataset with 100 bp reads from synthetic community with 64 species is used. Subfigure A shows a distance cutoff of 4% or 5% is proper for all regions sizes from 50 bp to 160 bp in V4 (OTU number approached the species number 64 as indicated by the black line). Different colors indicates region sizes in basepairs (bp). Subfigure B shows more details in the method used for subfigure A. Panel “Read length cutoff” in B shows minimum read length was set to the target region size minus 5 bp if the region size was less than 100 bp, and 95 bp when the region size was longer than 100 bp. As a result, the number of reads aligned decreased as the target region size increased until 100 bp, and then the number of reads aligned increased with target region as shown in Panel “Mapped read number” in B. Panel “OTU number” in C shows OTU number at distance cutoff of 0.05 and our method works well from a 50 bp region to 160 bp region in V4. Subfigure C tests our unsupervised method on multiple hyper-variable regions (V2, V3, V4, V5, V6, V8) with region size of 120 bp (circle) and 80 bp (triangle). Panel “Mapped read number” in C shows the number of reads mapped to each chosen region. Panel “OTU number” in C shows the number of OTUs in each region at distance cutoff of 0.05. All regions have consistent mapped read number and OTUs except V3.

Figure 2.3: Testing the effect of target region size on V4 of full-length SSU rRNA genes in clustering from a synthetic community with 64 species. The OTU number is close to species number (64) with all the region sizes. Different colors indicates region sizes in basepairs (bp).

Reproducibility between technical replicates is important and a basic feature of a method [78, 79]. We evaluated it by comparing the correlation of OTU abundance between technical replicates from the bulk soil sample and found high correlation between them (Pearson’s correlation coefficient = 0.997). The consistency was better for the more abundant OTUs (Fig. 2.4A). Log transformation could reduce the effect of high abundance OTUs on the overall correlation. Even with log-transformed abundance, the two replicates have a Pearson’s correlation coefficient.
of 0.91 and linear regression $R^2$ of 0.88 when OTUs with less than 25 total counts were discarded (Fig. 2.4B). The choice of 25 cutoff was chosen to compare to another study on amplicon reproducibility [80] and as mentioned in the Discussion.

Figure 2.4: Technical reproducibility test of our unsupervised clustering and comparison of OTU abundances between paired shotgun and amplicon data.
Figure 2.4 (cont’d): Subfigure A shows consistent OTU abundance profiles in two technical replicates (Pearson’s correlation coefficient is 0.997). X axis shows number of reads in each OTU in replicate SB1_123, and y axis shows number of reads in each OTU in replicate SB1_456. The size of circle is proportional to number of OTUs at the same location in the plot (with the same counts in SB1_123 and also in SB1_456). The consistency of counts of each OTU in two replicates becomes better when the abundance of OTUs are higher. Subfigure B shows progressive dropout analysis of two technical replicates of shotgun data. There is significant correlation of counts of each OTU between technical replicates; X axis is the threshold of OTU abundance and y axis is the $R^2$ of linear regression of log transformed OTU abundances in two replicates. OTUs with lower abundance than the thresholds (x axis) were discarded before regression analysis. Subfigure C and D shows comparison of OTU abundance profile between paired shotgun and amplicon data in bulk soil sample (SB1) and rhizosphere sample (M1), respectively. There is inconsistency between shotgun data and amplicon data in both samples. X axis shows number of SSU rRNA gene fragments in shotgun data per OTU in log scale, and y axis shows number of amplicon sequences in each OTU in log scale. The OTU abundance in both amplicon and shotgun data were increased by 1 to avoid 0 counts that can be displaced in log scale. The size of circle is proportional to number of OTUs with the same abundance in both types of data. There are OTUs with significantly different abundances in the two types of data (circles deviate from diagonal line). Pearson’s correlation between two types of data is 0.873 in SB1 and 0.581 in M1.

We also compared OTU-based microbial community structures inferred from shotgun and amplicon SSU rRNA gene sequences. OTU abundances in shotgun and amplicon data, however, do not correlate as well (Pearson’s correlation coefficient of 0.87 for the bulk soil sample SB1 (Fig. 2.4C) and 0.58 for the Miscanthus rhizosphere sample M1 (Fig. 2.4D), showing that the amplicon and shotgun methods are not providing the same information. The classification of OTUs with total abundance higher than 10 and with ratio between two data type higher than five fold shows Verrucomicrobia were biased against in the bulk soil sample (SB1_PT) amplified by V6-V8 primer, while they were biased to in the rhizosphere sample (M1_PT) amplified by V4 primer. Actinobacteria was biased against in M1_PT amplified by V4 primer (Fig. 2.5). The above results were consistent with the taxonomy-based comparison of the two data types (see below) that suggesting primer bias in amplicon data.

We applied ordination analysis to OTU tables from unsupervised analysis of corn and Miscanthus rhizosphere samples. OTUs from shotgun and amplicon data both showed separation of rhizosphere communities of corn and Miscanthus (horizontal dimension in Fig. 2.6), as well as a
Figure 2.5: **Phyla of OTUs significantly different between shotgun data and amplicon data.**

SB_123 are shotgun data and SB1_PT are amplicon data both from the same DNA from bulk soil sample. M1 are shotgun data and M1_PT are amplicon data both from the same DNA from Miscanthus rhizosphere sample. OTUs significantly different were defined as those with total abundance > 10 and fold change between two types of data > 5 or < 0.2. Verrucomicrobia was biased against in bulk soil sample amplified by V6-V8 primer (SB1_PT) but biased for in rhizosphere sample amplified with V4 primer (M1_PT). Actinobacteria was biased against in rhizosphere sample (M1).

Significant difference between the two data types (vertical dimension in Fig. 2.6), confirming the difference between shotgun and amplicon data. Significant separation (p < 0.001 by AMOVA test in Mothur) of corn and Miscanthus samples was also observed when V2, V4, V6, and V8 shotgun data were used for clustering (Fig. 2.7) but the sample groupings were the same for all variable regions. Figures 2.6 and 2.7 showed that the dispersion among the seven corn replicates was much higher than for the Miscanthus replicates. Miscanthus samples had higher alpha diversity than corn samples as shown for all of V2, V4, V6 and V8 regions, although there were variations among these regions (Fig. 2.8).

We compared the taxonomy-based microbial community structures inferred from shotgun data with those from amplicon SSU rRNA gene sequences (12163 amplicon for SB1 and 60148 for M1) and confirmed known primer biases and revealed a new bias. Before comparing two data types, we looked at the taxonomy profile of shotgun data using different variable regions. Shotgun data mapped to different variable regions shows similar taxonomy at the Bacterial phylum level.
Figure 2.6: **Principle coordinates analysis (PCoA) of amplicon and shotgun derived data from seven field replicates.** There are significant differences between amplicon (filled) and shotgun (unfilled) derived data (along y-axis) and between corn (circle) and Miscanthus (square) rhizosphere samples (along x-axis), (AMOVA p-value < 0.001). PCoA was applied to OTU table resulting from clustering with shotgun data and amplicon data using 150bp of V4 region. Labels with suffix “_PT” are amplicon data and others are shotgun data. (Pearson’s correlations > 0.96) except that V6 has more unclassified sequences (Fig. 2.9). Since different variable regions may provide different taxonomic precision to certain groups [81], taxonomy information from all regions may better represent the taxonomy profile. Thus we used all SSU rRNA gene fragments for taxonomy comparison with amplicon data. Both shotgun and amplicon data show Actinobacteria, Proteobacteria and Acidobacteria as the three most abundant phyla, as is expected for soil [82]. Since shotgun data are more accurate on estimating community structure, thus we accept the shotgun data as the reference [5, 58]. The 926F/1392R (V6-V8) primer set is biased against Verrucomicrobia (0.3% in amplicon data vs. 5.8% in shotgun data by RDP
Figure 2.7: **Comparison of ordination analysis with different variable regions.** PCoA of OTUs from different SSU rRNA variable regions (V2, V4, V6, and V8) was applied on corn (filled markers, “_C”) and Miscanthus (unfilled markers, “_M”) rhizosphere samples. Different colors indicate the seven replicates. OTU tables from clustering of shotgun data using 150bp of V2, V4, V6 and V8 regions were used for PCoA and procrustes analysis in QIIME was used to transform the PCoA results from different regions and plot them in the same figure.

database) in bulk soil sample SB1 (Fig. 2.10), and the 515F/806R (V4) primer set biased against Actinobacteria (11.6% in amplicon data vs. 26.6% in shotgun data) and towards Verrucomicrobia (5.9% in amplicon data vs. 3.2% in shotgun data) in rhizosphere sample M1 (Fig. 2.10).

To take advantage of the fact that shotgun data is un-targeted, we retrieved and classified the Large Subunit (LSU) rRNA genes, which are co-transcribed with SSU rRNA genes. Their taxonomy profile was similar to that of SSU rRNA gene (Pearson’s correlation coefficient of 0.87 for SB1 and 0.91 for M1), except that more reads (19.6%) remain unclassified (Fig. 2.10). This is expected because of the much lower number of reference LSU rRNA genes in the SILVA database.
Figure 2.8: **Alpha diversity comparisons between corn and Miscanthus using V2, V4, V6 and V8 regions.** All variable regions showed Miscanthus is more diverse than corn, even though there was variation of diversity among rRNA gene regions. Alpha diversity was calculated with inverse Simpson using OTUs resulting from clustering with different variable regions.

The two genes show consistent community profiles at the Bacterial phylum level and also confirm the known primer bias against Verrucomicrobia in 926F/1392R (V6-V8) primer set and also the bias against Actinobacteria in 515F/806R (V4) primer set (Fig. 2.10). Further, both the LSU and SSU HMMs show the ability to identify Eukaryota members and give about the same taxonomy profile at domain level (Fig. 2.11). It is noteworthy that both LSU and SSU shotgun data show Archaea to be twice as numerous (2% vs. 1%) in the bulk soil as the rhizosphere and that Eukaryota were much more numerous (6% vs. 1%) in the Miscanthus rhizosphere soil than the bulk soil (Fig. 2.11). Higher fungal percentage (2.54% vs. 0.39%), fungi/bacteria ratio (0.027 vs. 0.004), and Arbuscular Mycorrhizal Fungi (AMF) percentage in fungi (0.18% vs. 0) are found in
Figure 2.9: **Bacterial phylum profile comparison using different variable regions.** Different variable regions have similar taxonomy profiles, except that V6 has more unclassified sequences. The minimum Pearson’s correlation between the regions is 0.96. Classifications were done using SSU rRNA gene fragments from Miscanthus rhizosphere soil sample (M1) and SILVA database as reference.

rhizosphere sample (M1) than bulk soil sample (SB1) (Table 2.2). Copy correction was applied on two soil samples (SB1 and M1). Both samples showed that Firmicutes and Bacteroidetes had the highest fold change after copy number correction (Fig. 2.12). Despite copy number correction, the clustering of our soil samples did not change (Fig. 2.13) compared to that without copy number correction (Fig. 2.6), probably because of the low proportion of taxa with large rrn number corrections.
Figure 2.10: Taxonomy profiles of Bacterial phyla of shotgun fragments from SSU and LSU rRNA genes and of amplicon reads. Classifications were done using both RDP and SILVA reference databases. “_PT” indicates amplicon data. Lower panel is from bulk soil sample (SB1) and its amplicon data (_PT) uses V6-V8 primer (515F/806R) and shows less Verrucomicrobia detected using both databases. Upper panel is rhizosphere sample (M1) and its amplicon data uses V4 primer (515F/806R) and shows less Actinobacteria and more Verrucomicrobia using both databases.
Figure 2.11: **Taxonomy profile comparison at domain level using SSU and LSU rRNA genes.** For both the bulk soil sample (SB1) and rhizosphere sample (M1), SSU and LSU show consistent domain level taxonomy distribution (Pearson’s correlation coefficient = 1). “_LSU” indicates taxonomy from LSU rRNA SILVA database and the rest are classified by SSU rRNA database.
Figure 2.12: Bacterial phylum level taxonomy summary before and after SSU rRNA gene copy correction.
Figure 2.12 (cont’d): Left vertical axis with bar plot shows percentage in total community, while right vertical axis with line plot shows fold change after copy number correction. Taxa with relative abundances of more than 0.1% before copy correction were chosen and were ordered based on fold change. Subfigure A is for bulk soil sample (SB1) and B is for Miscanthus rhizosphere sample (M1).

Table 2.2: **Higher fungi/bacteria ratio and percent of AMF fungi are in rhizosphere sample (M1) than bulk sample (SB1).**

<table>
<thead>
<tr>
<th></th>
<th>% Bacteria</th>
<th>% Fungi</th>
<th>% AMF in Fungi</th>
<th>Fungi/Bacteria</th>
</tr>
</thead>
<tbody>
<tr>
<td>SB1-SSU</td>
<td>97.00%</td>
<td>0.36%</td>
<td>0.00%</td>
<td>0.0037</td>
</tr>
<tr>
<td>SB1-LSU</td>
<td>96.75%</td>
<td>0.42%</td>
<td>0.00%</td>
<td>0.0044</td>
</tr>
<tr>
<td>M1-SSU</td>
<td>92.38%</td>
<td>2.60%</td>
<td>0.18%</td>
<td>0.0281</td>
</tr>
<tr>
<td>M1-LSU</td>
<td>92.94%</td>
<td>2.48%</td>
<td>0.18%</td>
<td>0.0267</td>
</tr>
</tbody>
</table>

2.5 Discussion

We present, characterize and validate an efficient method for retrieving and analyzing SSU rRNA gene fragments from shotgun metagenomic sequences. The pipeline enables unsupervised diversity analysis with copy number correction on multiple variable regions, has the scalability to handle large soil metagenomes, is expandable to other phylogenetic marker genes and is publicly available on GitHub.

We apply a two-step approach for retrieving SSU rRNA gene fragments, a loose HMM filtering step followed by a more stringent step that screens by identity to the best match reference. The first step leverages HMMER [39] and should thus have better scalability than existing shotgun analysis pipelines that use BLAST-like tools such as MG-RAST [83]. MG-RAST annotates shotgun reads by BLAT search against rRNA databases and the taxonomy of reads is inferred from the best hit or least common ancestor of several top hits [31, 65, 84]. BLAT or BLAST-like tools are not scalable for large datasets and must be run in parallel on large computer clusters, because BLAST-like tools typically do pairwise comparison of reads against large and growing rRNA database such
Figure 2.13: Ordination plot of amplicon and shotgun derived data after copy correction. The plot is similar to the one without copy correction (Fig. 2.8). There were significant differences in amplicon and shotgun derived data (y-axis) and of corn and Miscanthus rhizosphere samples (x-axis), (AMOVA p-value < 0.001), after copy number correction. PCoA was applied to OTU table resulting from de novo clustering with shotgun data and amplicon data using 150bp of V4 region. The filled markers (“_PT”) are amplicon data and the unfilled markers are shotgun data.

as RDP, SILVA and GreenGenes [23, 24, 76], while HMM-based methods compare reads to only fixed number of models (commonly one for each domain) and thus more scalable [83]. Moreover, these pipelines lack unsupervised community analysis. HMM-based search has been used before as it is fast and sensitive for rrn retrieval [33, 60, 61, 63] and current existing implementations such as meta-rna, RNASelect and metaxa are all wrappers around HMMER [39]. We chose meta-rna and metaxa for comparison, for the reason that RNASelect can only run in graphic interface that is not suitable for large datasets.

The second step evaluates hmmsearch results based on identity to their best-match reference
and also prepares the alignment of SSU RNA gene fragments for clustering. Since there is no clear sequence identity threshold for SSU rRNA genes, the choice of identity cutoff is arbitrary (the default is 50%). This is also common practice in amplicon analysis platforms where reads with low identity to reference sequences are discarded prior to clustering [34, 35]. For consistency in comparison, sequences in our amplicon datasets with less than 50% identity to reference sequences are also discarded. An alignment of the SSU rRNA gene fragments is essential for the later unsupervised analyses. In comparison to methods that use only the 16S rRNA gene in E.coli as alignment template [85], our method takes advantages of the rich phylogenetic diversity of SSU rRNA genes provided by the SILVA database. Increasing the number of reference sequences can improve the quality of alignment, but also linearly increases the memory required [72, 86].

Our unsupervised analysis with shotgun data is a novel and important part of the pipeline. Our tests shows that regions as small as 50 bp could be applied to clustering. Thus short reads around 50 bp could be applied to this method as long as there are sufficient numbers of reads aligned to the target region (sequencing depth is a limiting factor; see below). This is consistent with pilot studies from 454 amplicon sequencing [30, 87]. When sequencing depth is limited, there is flexibility to control the number of reads to include for clustering by adjusting the target region size and read length cutoff within certain limits (Fig. 2.2 B). The caveat of using very short or very large target regions is that the overlapping portion of reads will decrease and thus decrease the accuracy of clustering and the impact of sequencing error will increase; for example, an error in a 50 bp read can cause 2% distance and we will accordingly need to set a larger distance cutoff for clustering. In addition, we can obtain longer sequences from overlapping paired ends as shown in our soil data (Fig. 2.14). Thus reads from Illumina shotgun data (ranging from 75 to 250 bp) can be used for unsupervised analysis. Note that the flexibility on choice of variable region for analysis is another advantage of shotgun data (Fig. 2.2 C).

We also found good reproducibility of OTU abundance between technical replicates, which is critical for the validity of our method (Fig. 2.4 A). Generally, OTU-based analysis provides higher resolution than taxonomy-based diversity analysis for community comparison, largely due to the
Figure 2.14: **Length distribution of trimmed reads after quality trimming and paired-end merging.** SB1 is the bulk soil data and M1 is the rhizosphere data. The reads with >150 bp result from the merged paired ends, which benefits classification and clustering in downstream analyses. Reads less than 150 bp are also used in the analysis and come from unmerged paired reads.

Databases lacking reference sequences from uncultured microbes [88]. The high correlation of OTU abundance in two technical replicates shows the reproducibility of the analysis of shotgun data, which is comparable to the reproducibility of amplicon data shown in another study in terms of Pearson’s correlation and $R^2$ of linear regression [80]. Further, comparison of OTU abundances in shotgun data and amplicon data sequenced from the same DNA extraction also show that many OTUs have inconsistent abundances between the two types of data (Fig. 2.4 C and D), which agrees with the differences seen in the taxonomy-based comparison (Fig. 2.10).

Community comparison by ordination methods such as PCoA and NMDS is one of the most common analyses in microbial ecology. To the best of our knowledge, the methods used in two studies [85, 89] are the only existing tools that are designed to deal with clustering of SSU rRNA gene fragments from Illumina shotgun data. The former could result in poor alignment by using only the 16S rRNA gene in E. coli as the alignment template. The latter (PhylOTU) was applied to longer shotgun sequences from Sanger sequencing in their study. It does the OTU clustering of SSU rRNA gene fragments aligned together over the whole gene length, which can be problematic.
because fragments aligned to different regions do not overlap and thus the clustering results are not reliable, even though the reference sequences included in the clustering process can improve the results. Since our tests shows that as low as 50 bp hyper-variable region can be used for unsupervised analyses (Fig. 2.2 A and C), we make sure all the sequences overlap by picking one small region (150 bp) and all sequences included in clustering have lengths longer than 100 bp in our clustering method. In addition, longer reads obtained by assembling overlapping paired end reads (Fig. 2.14) can make use of more overlap among reads and thus are more suitable for clustering. As read lengths increase with improvement of sequencing technology, longer regions can be chosen and the clustering results will be even more reliable. Also, shotgun data provides the flexibility to choose any variable region (Figs. 2.2 and 2.7 to 2.9), and the consistency of results from different variable regions provides more confidence in the biological conclusion as well as the method itself.

Primer bias is a major limitation of amplicon methods, which was apparent in our comparisons of community profiles from amplicon versus shotgun data (Fig. 2.10). Commonly, it is difficult to tell a bias is caused by primer or DNA extraction. The paired amplicon and shotgun data from the same DNA extract provides us a new opportunity to evaluate this issue, since the difference is only from the sequencing step. We used two main SSU rRNA gene databases, RDP and SILVA, to make sure the taxonomy distribution was not biased by the choice of reference databases and further confirmed by taxonomy distribution by the LSU rRNA gene. The bias against Verrucomicrobia with V6-V8 primers is consistent with other studies showing that Verrucomicrobia’s abundance in soil samples is underestimated due to primer bias [56]. Meanwhile, the bias towards Verrucomicrobia with V4 primers agrees with studies showing that V4 primer set has better coverage of Verrucomicrobia [56]. Furthermore, the V4 primer set shows bias against Actinobacteria. The V4 primer set is reported to cover 92.4% of Actinobacteria in reference databases, the lowest among nine common bacterial phyla [56]. Bias against Actinobacteria has also been reported in a study on synthetic community where no primer mismatch was found with members from Actinobacteria [58], and another study on environmental samples using Sanger sequencing (24F/1492R) [90], suggesting that
in-silico evaluation of primers is not sufficient and some factors other than primer mismatch are causing the bias, for example, competition between primers and melting temperature [91]. Thus primer bias detection by comparing paired amplicon and shotgun data is superior to methods that only search primers in the reference databases.

Another advantage of our method over amplicon approaches is that we can identify the SSU rRNA gene from Bacteria, Archaea and Eukaryota. Fungi are of special interest in microbial ecology due their critical role in ecosystems [92, 93]. The Fungi/Bacteria ratio is an important indicator of C/N ratio and soil health [94, 95]. Our shotgun metagenome (DNA based) shows a ratio of 0.004 in bulk soil (SB1) and 0.027 in rhizosphere soil (M1) (Table 2.2), while studies using Phospholipid Fatty-acid Analysis (PLFA) at the same sampling site (KBS) typically shows a ratio of 1 to 1.3 (58). The difference can be explained by higher biomass of fungi relative to DNA, since some fungi hyphae may not be filled with nuclei. In addition, we also found higher percentage of AMF (Abuscular Mycorrhiza Fungi) of rhizosphere soil (M1) compared to bulk soil (SB1), which is consistent with their symbiotic relationship with grass roots.

We also show that copy number correction can be achieved in our pipeline. Gene copy number is another source of bias that limits one’s ability to accurately profile microbial communities. There are up to 15 SSU rRNA gene copies in some Bacteria and up to 5 in Archaea [96]. This pipeline utilizes the SSU rRNA copy database in CopyRighter [75]. As expected, the Firmicutes in soil samples have the highest fold change (Fig. 2.12). But, due to their low proportion in these soils, the impact of copy number correction on the overall community profile was minor (Fig. 2.13). Copy number correction, however, is still an open problem because SSU rRNA gene copy number data for most species and/or OTUs are lacking and copy number can be incorrect even for species with complete genome sequences because of mis-assembly of these repeated regions.

Sequencing depth is another important factor in considering this method for diversity analysis. The percentage of SSU rRNA gene fragments in shotgun data varies depending on the SSU rRNA gene copy number and genome size of each member. In our bulk soil sample (SB1) and the Miscanthus rhizosphere soil sample (M1), we classified about 0.03% and 0.04% of the total
shotgun data as SSU rRNA, respectively. In an ideal situation we want to get enough SSU rRNA gene fragments to see saturation of rarefaction curve in OTU-based analysis, which is difficult for soil samples because of their high diversity and the presence of sequencing error. However, studies have shown that near saturation sequencing of SSU rRNA amplicons is not necessary for beta-diversity analysis [54, 97]. Thus empirical fold coverage of 3000 based on the whole length (about 1500 bp) of SSU rRNA gene is suggested for surface soil samples, which requires 11.2 (1500*3000/0.04%) Gbp of shotgun data, assuming the SSU rRNA gene comprises about 0.04% of total data.

In this study, the LSU rRNA gene was used mainly as confirmation for SSU rRNA gene-based diversity analysis (Figs. 2.10 and 2.11). However, the LSU rRNA gene offers additional stretches of variable and characteristic sequence regions due to its longer sequence length, and thus yields better phylogenetic resolution [98]. For the above reason and also the reason that there are more available references for fungi, the LSU rRNA gene is more commonly used for fungal community studies [99–102]. Currently the use of the LSU rRNA gene is limited by reference sequences and universal primer sets available [98], but its increased resolution should not be overlooked since too limited resolution of the SSU rRNA gene is a barrier in many ecological studies [92]. In the future, other single copy genes with phylogenetic reference and finer resolution, such as rplB, gyrB, and recA, could also be recovered from metagenomic sequence and used for community structure analysis by a similar pipeline [41].

### 2.6 Conclusion

We developed a fast and efficient pipeline that enables unsupervised diversity analysis with Illumina shotgun data. The pipeline has the scalability to analyze large datasets (5 CPU hours for 38 Gb data, with 4.8 Gb peak memory) and can be run on most desktops with more than 5 GB of memory. Since SSU rRNA based community analysis is an important method in microbial ecology, this method can save projects with existing shotgun sequence data from the additional cost of SSU rRNA amplicon sequencing. Moreover, shotgun sequencing are not as affected by primer bias
and chimeras as amplicon sequencing and thus can improve the measurement of microbial community structure. As read length and sequencing depth increases, longer and more SSU rRNA gene fragments can be recovered. Thus clustering and diversity analysis by this pipeline will become even more reliable.
CHAPTER 3
GENOMIC AND METAGENOMIC COMPARISON OF rplB AND SSU rRNA GENE IN SPECIES RESOLUTION

3.1 Introduction

Shaped by 3.5 billion years of evolution, microorganisms are estimated to comprise up to one trillion species and the majority of genetic diversity in the biosphere [103]. However, our understanding of this diversity is limited, mainly because the majority of microorganisms cannot be cultured under laboratory conditions. Since the pioneering work of Carl Woese in the late 1970s, the SSU rRNA gene has become the dominant marker used in in microbial community diversity analysis [21, 22, 54, 104]. The fact that the SSU rRNA gene is highly conserved and that there are often multiple copies in the same genome with intra-genomic variations makes the SSU rRNA gene problematic for taxonomic identification at species and strain levels [40, 41, 105].

With the accelerated accumulation of microbial genomes in NCBI in recent years [106], whole genome based comparison is considered to be the most accurate method for species and strain identification [106–110]. However, whole genome based comparison is computationally more expensive compared to marker gene comparison, and more importantly, we are unable to reliably obtain genome sequences of members in a microbial community, especially for complex environments such as soil. Thus, genome based taxonomic definitions are less useful for community diversity analysis, where marker gene analysis is still essential.

Because of the disadvantages of using the SSU rRNA gene, microbial ecologists have been looking for alternatives. Single copy protein coding housekeeping genes stand out for the following reasons. First, they are single copy in the genome and thus allow more accurate species and strain identification and OTU clustering than the SSU rRNA gene which is limited by intra-genomic heterogeneity. Second, they are present in virtually all members of the three domains of life. Third, protein coding genes evolve faster than rRNA genes not only because rRNA genes are more conserved due to their critical role in ribosomes [111], but also because of the redundancy in the
genetic code, especially at the third codon position [40].

Here, we use a single copy protein coding gene, \textit{rplB} (coding ribosomal large subunit protein L2) to showcase the potential of using single copy housekeeping genes as a phylogenetic marker for microbial community diversity analysis. Although there are studies using either genomic data or metagenomic data [40,41], we compare \textit{rplB} and SSU rRNA genes with all completed bacterial genomes (a genome as a single sequence) and also large soil shotgun metagenomic data that are orders of magnitude larger than those used in previous studies. The novelty of our genomic data analyses is that we evaluate difference in quantity (i.e. percent identities among taxons in gene sequence) rather than whether taxons are different in gene sequence). In metagenomic data analyses, we focus on the different effects of the two different marker genes on de novo OTU based analyses (a common practice in microbial diversity analyses), while previous studies focus on taxonomic identification [40,41].

3.2 Methods

3.2.1 Downloading bacterial genomes

Bacterial genome assembly information from NCBI (ftp://ftp.ncbi.nlm.nih.gov/genomes/refseq/bacteria/assembly_summary.txt) was used to construct the url to download each genome based on the instructions described in this link (http://www.ncbi.nlm.nih.gov/genome/doc/ftpfaq/#allcomplete). Command line “wget” was then used to retrieve the genome sequences with urls obtained from the above step.

3.2.2 SSU rRNA gene and \textit{rplB} extraction and pairwise comparison

The SSU rRNA gene HMM (Hidden Mokov Model) from SSUsearch was used [6]. Aligned \textit{rplB} nucleotide sequences of the “training set” retrieved from the RDP FunGene database were used to build HMM using hmmbuild command in HMMER (version 3.1b2) [39,112]. The nhmmer command in HMMER was then used to extract SSU rRNA gene sequences and \textit{rplB} sequences
from the Bacterial genomes obtained from NCBI using score cutoff (-T) of 60. Next, nhmmer at least 90% as long as the HMM model were accepted as the correct target gene. For the purpose of comparing SSU rRNA gene distance and rplB gene distance, one copy of the SSU rRNA gene was randomly picked for each genome. Pairwise comparison among gene sequences was done using vsearch (version 1.1.3) [113] with “–allpairs_global –acceptall”. Three species of interest: *Rhizobium leguminosarum*, *Pseudomonas putida* and *Escherichia coli*, were chosen for closer comparison of rplB pairwise distances and SSU rRNA gene distances.

3.2.3 Rhizosphere soil metagenomic data download and preprocess

Shotgun sequencing metagenomic data for 21 samples were downloaded from the JGI web portal ([http://genome.jgi.doe.gov/](http://genome.jgi.doe.gov/)). JGI Project IDs for the 21 samples are listed in Table 3.1. The samples are from the rhizosphere of three biofuel crops: corn (C), switchgrass (S) and *Miscanthus* (M), with seven field replicates for each crop. Raw reads were quality trimmed using fastq-mcf (version 1.04.662) ([http://code.google.com/p/ea-utils](http://code.google.com/p/ea-utils)) “-l 50 -q 30 -w 4 -k 0 -x 0 –max-ns 0 -X”. Overlapping paired-end reads were merged by FLASH (version 1.2.7) [71] with “-m 10 -M 120 -x 0.20 -r 140 -f 250 -s 25 -t 1”, which has also been described in [6].

3.2.4 SSU rRNA gene and rplB identification and comparison in soil metagenomes

SSU rRNA gene fragments and those aligned to the V4 region (*E. coli* position: 577 - 727) of each sample were identified using the SSUsearch pipeline [6] and clustered using the RDP’s McClust tool [23]. In order to evaluate the effect of the two genes on OTU number, SSU rRNA gene reads identified using SSUsearch were assembled into full length SSU rRNA gene sequences using emirge_amplicon.py from the EMIRGE package (version 0.60.3) [114] with options “-l 150 –phred 33 -a 4 -i 275 -s 25”, to compare with full length rplB sequences described below on OTU number.

Full length rplB sequences were assembled using Xander with flags “MAX_JVM_HEAP=500G, FILTER_SIZE=40, K_SIZE=45, genes=rplB, THREADS=9” [36]. Data for each crop were assembled separately. The assembled rplB sequences (nucleotide and
Table 3.1: Rhizosphere soil shotgun metagenomic data.

<table>
<thead>
<tr>
<th>Sample</th>
<th>JGI project ID</th>
<th>Data size (Gb)</th>
</tr>
</thead>
<tbody>
<tr>
<td>C1</td>
<td>1023764</td>
<td>46</td>
</tr>
<tr>
<td>C2</td>
<td>1023767</td>
<td>39</td>
</tr>
<tr>
<td>C3</td>
<td>1023770</td>
<td>57</td>
</tr>
<tr>
<td>C4</td>
<td>1023773</td>
<td>53</td>
</tr>
<tr>
<td>C5</td>
<td>1023776</td>
<td>57</td>
</tr>
<tr>
<td>C6</td>
<td>1023779</td>
<td>51</td>
</tr>
<tr>
<td>C7</td>
<td>1023782</td>
<td>46</td>
</tr>
<tr>
<td>M1</td>
<td>1023785</td>
<td>56</td>
</tr>
<tr>
<td>M2</td>
<td>1023788</td>
<td>57</td>
</tr>
<tr>
<td>M3</td>
<td>1023791</td>
<td>50</td>
</tr>
<tr>
<td>M4</td>
<td>1023794</td>
<td>42</td>
</tr>
<tr>
<td>M5</td>
<td>1023797</td>
<td>45</td>
</tr>
<tr>
<td>M6</td>
<td>1023800</td>
<td>43</td>
</tr>
<tr>
<td>M7</td>
<td>1018623, 1018611</td>
<td>32</td>
</tr>
<tr>
<td>S1</td>
<td>1023803</td>
<td>40</td>
</tr>
<tr>
<td>S2</td>
<td>1018626, 1018614</td>
<td>28</td>
</tr>
<tr>
<td>S3</td>
<td>1023806</td>
<td>44</td>
</tr>
<tr>
<td>S4</td>
<td>1023809</td>
<td>49</td>
</tr>
<tr>
<td>S5</td>
<td>1018629, 1018617</td>
<td>27</td>
</tr>
<tr>
<td>S6</td>
<td>1023812</td>
<td>50</td>
</tr>
<tr>
<td>S7</td>
<td>1023815</td>
<td>39</td>
</tr>
</tbody>
</table>

protein) from three crops were pooled and clustered using RDP’s McClust tool [23]. An OTU table for each gene with OTU counts of each sample were made based on mean kmer coverage of the representative sequence of each OTU.

Further, diversity analyses were done by vegan package in R using functions “rda” for ordination and “diversity” for Shannon diversity index, respectively, from the OTU (count) tables. For comparing the resolution of two genes assembled from the metagenomic data, since the SSU rRNA gene had higher sampling depth than \( rplB \) due to the multiple copy nature of SSU rRNA genes, we re-inflated the unique full length sequences of both genes based on the abundance information (i.e. to multiply each unique sequence by its abundance) and then we evenly subsampled the two gene datasets. The abundance information of \( rplB \) sequences was in a file with the name “*_coverage.txt” in the \( rplB/cluster \) directory from Xander. Meanwhile, the relative abundance information of SSU rRNA gene assemblies can be found the description field of assembled sequence file in
fasta format as described in README.txt (NormPrior) of the EMIRGE package. With the total coverage of the gene estimated from the reads identified as SSU rRNA gene sequences (total_bp / 1500), the abundance of each gene was estimated by multiplying relative abundance and total coverage of the SSU rRNA gene.

### 3.3 Results

#### 3.3.1 SSU rRNA gene and rplB copy number and intra-genome SSU identity

A total of 53865 Bacterial genomes were downloaded with 4457 of them complete. The maximum and mean of SSU rRNA gene copy number detected was 20 and 2 among all genomes, and 16 and 4 among complete genomes (Table 3.2). Multiple copies of rplB within the same genome were also detected (Table 3.2), with a maximum of 6 among all genomes and 2 among complete genomes. In order to compare these two genes, they need to both be present in genomes. We found 46600 genomes and 4440 complete genomes with both genes. When looking at intra-genomic variation among copies in completed genomes of *R. leguminosarum*, *P. putida* and *E. coli*, we found that *E. coli* has the largest variation among the three, with a minimum of 95.4% (Fig. 3.1). For the pairwise comparison of genomes below, one copy of SSU rRNA gene was randomly picked as a representative for genomes with multiple copies.

![Figure 3.1](image.jpg)

Figure 3.1: Intra-genome variation of SSU rRNA gene of *R. leguminosarum*, *P. putida*, and *E. coli* in boxplot. *E. coli* has largest variation among SSU rRNA gene copies within the three. Circles are outlies in boxplots. Each dot is a comparison between two copies of SSU rRNA gene in the same genome.
Table 3.2: Summary of SSU rRNA gene and rplB copy number in genomes.

<table>
<thead>
<tr>
<th>Copy</th>
<th>SSU_all</th>
<th>SSU_complete</th>
<th>rplB_all</th>
<th>rplB_complete</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>32576</td>
<td>635</td>
<td>53289</td>
<td>4443</td>
</tr>
<tr>
<td>2</td>
<td>3136</td>
<td>711</td>
<td>90</td>
<td>4</td>
</tr>
<tr>
<td>3</td>
<td>2649</td>
<td>560</td>
<td>14</td>
<td>NA</td>
</tr>
<tr>
<td>4</td>
<td>2264</td>
<td>585</td>
<td>3</td>
<td>NA</td>
</tr>
<tr>
<td>5</td>
<td>2006</td>
<td>372</td>
<td>NA</td>
<td>NA</td>
</tr>
<tr>
<td>6</td>
<td>1603</td>
<td>484</td>
<td>1</td>
<td>NA</td>
</tr>
<tr>
<td>7</td>
<td>1333</td>
<td>556</td>
<td>NA</td>
<td>NA</td>
</tr>
<tr>
<td>8</td>
<td>525</td>
<td>194</td>
<td>NA</td>
<td>NA</td>
</tr>
<tr>
<td>9</td>
<td>222</td>
<td>92</td>
<td>NA</td>
<td>NA</td>
</tr>
<tr>
<td>10</td>
<td>187</td>
<td>107</td>
<td>NA</td>
<td>NA</td>
</tr>
<tr>
<td>11</td>
<td>114</td>
<td>58</td>
<td>NA</td>
<td>NA</td>
</tr>
<tr>
<td>12</td>
<td>72</td>
<td>27</td>
<td>NA</td>
<td>NA</td>
</tr>
<tr>
<td>13</td>
<td>58</td>
<td>26</td>
<td>NA</td>
<td>NA</td>
</tr>
<tr>
<td>14</td>
<td>48</td>
<td>38</td>
<td>NA</td>
<td>NA</td>
</tr>
<tr>
<td>15</td>
<td>8</td>
<td>4</td>
<td>NA</td>
<td>NA</td>
</tr>
<tr>
<td>16</td>
<td>7</td>
<td>1</td>
<td>NA</td>
<td>NA</td>
</tr>
<tr>
<td>18</td>
<td>1</td>
<td>NA</td>
<td>NA</td>
<td>NA</td>
</tr>
<tr>
<td>20</td>
<td>1</td>
<td>NA</td>
<td>NA</td>
<td>NA</td>
</tr>
</tbody>
</table>

3.3.2 Pairwise comparison of genomes with SSU rRNA gene and rplB

Overall, rplB gene sequences were less similar than SSU rRNA gene sequence for most genome pairs (Fig. 3.2). SSU rRNA gene sequence pairwise similarities were mainly between 90% and 100%, while rplB distances were in a larger range between 70% and 100%. Further, we targeted several species of interest for detailed examination: R. leguminosarum, P. putida and E. coli and found all selected species except E. coli rplB genes less similar than SSU rRNA genes within their corresponding family, genus, and species (among strains) (Figs. 3.3 to 3.5). In contrast, most comparisons within genus and species for E. coli showed the SSU rRNA genes less similar than the corresponding rplB genes. SSU rRNA gene similarities were mostly 99% while most rplB genes were identical (Fig. 3.5 B and C). The above results were from complete genomes. We also applied the same analyses with all genomes but some strains within the E. coli showed low identities (< 80%) in their SSU rRNA genes (Fig. 3.6), indicating poor genome quality or strain mis-identification. Thus we decided to use only results from completed genomes only.
Figure 3.2: SSU rRNA gene and rplB identities among species in the same genus. X axis is SSU rRNA gene identities and Y axis is rplB identities. The dashed line is y = x. The histograms on the top and right axes are identity distributions for SSU rRNA gene and rplB respectively. SSU rRNA gene identities are larger than rplB in most genomes (most dots are below the dashed line), suggesting rplB has more variation among species in the same genus.
Figure 3.3: **SSU rRNA gene and rplB among R. leguminosarum related genomes.** SSU rRNA gene identities are larger than rplB in most genome pairs. X axis is SSU rRNA gene identities and Y axis is rplB identities. The dashed line is y = x. Size of dots indicates number of genome pairs that share the same SSU rRNA gene identity and rplB identity. A. All completed genomes in Rhizobiales are included in pairwise comparison; B. All completed genomes in *Rhizobium* are included; C. All completed genomes in *R. leguminosarum* are included.
Figure 3.4: **SSU rRNA gene and rplB among P. putida related genomes.** SSU rRNA gene identities are larger than rplB in most genome pairs. X axis is SSU rRNA gene identities and Y axis is rplB identities. The dashed line is y = x. Size of dots indicates number of genome pairs that share the same SSU rRNA gene identity and rplB identity. A. All completed genomes in Pseudomonadales are included in pairwise comparison; B. All completed genomes in *Pseudomonas* are included; C. All completed genomes in *P. putida* are included.
Figure 3.5: **SSU rRNA gene and rplB identities among E. coli related genomes.** X axis is SSU rRNA gene identities and Y axis is rplB identities. The dashed line is y = x. Size of dots indicates number of genome pairs that share the same SSU rRNA gene identity and rplB identity. A. All completed genomes in Enterobacteriales are included in pairwise comparison; B. All completed genomes in *Escherichia* are included; C. All completed genomes in *E. coli* are included.
Figure 3.6: **SSU rRNA gene and rplB identities among all E. coli genomes.** SSU rRNA gene identities are smaller than rplB in most genome pairs and some are as low as 70%. X axis is SSU rRNA gene identities and Y axis is rplB identities. The dashed line is y = x. Size of dots indicates number of genome pairs that share the same SSU rRNA gene identity and rplB identity.

### 3.3.3 Comparison of SSU rRNA gene and rplB on diversity analyses in large soil shotgun metagenomes

On average, 0.04% of total reads were identified as SSU rRNA gene fragments and 0.004% of total reads aligned to the V4 region of the gene with SSUsearch. Another 0.01% of total reads were identified as rplB with Xander (Table 3.3). Further, we found that the SSU rRNA gene had more OTUs than rplB without evenly subsampling with a distance cutoff range of 0 to 10% (Fig. 3.7 lower panel). However, after evenly subsampling sequences to 6000 for each gene, we observed rplB nucleotide sequences had more OTUs than SSU rRNA gene sequences consistently, and rplB protein sequences had more OTUs than SSU rRNA gene sequences except for distance offs ranging from 2% to 5% but (Fig. 3.7 upper panel). Regardless of the differences in OTUs resolution, both genes produced the same biological conclusions in two common diversity analyses: the microbial communities in rhizosphere soil of corn were different from those in Miscanthus and switchgrass (beta diversity by ordination) (Fig. 3.8); and corn had a significantly lower diversity than Miscanthus and switchgrass (alpha diversity) (Fig. 3.9). In addition, rplB also separated the
microbial communities in *Miscanthus* and switchgrass while SSU rRNA gene did not, suggesting higher resolution of the *rplB* genes (Fig. 3.8).

### Table 3.3: Summary of reads identified as SSU rRNA gene and *rplB* in soil metagenomes.

<table>
<thead>
<tr>
<th></th>
<th>TotalReads</th>
<th>SSU</th>
<th>SSU %</th>
<th>V4</th>
<th>V4 %</th>
<th>rplB</th>
<th>rplB %</th>
</tr>
</thead>
<tbody>
<tr>
<td>C1</td>
<td>232234095</td>
<td>117160</td>
<td>0.050%</td>
<td>11615</td>
<td>0.005%</td>
<td>29743</td>
<td>0.013%</td>
</tr>
<tr>
<td>C2</td>
<td>220844998</td>
<td>95493</td>
<td>0.043%</td>
<td>8137</td>
<td>0.004%</td>
<td>22605</td>
<td>0.010%</td>
</tr>
<tr>
<td>C3</td>
<td>282123335</td>
<td>119210</td>
<td>0.042%</td>
<td>11579</td>
<td>0.004%</td>
<td>33076</td>
<td>0.012%</td>
</tr>
<tr>
<td>C4</td>
<td>260542662</td>
<td>107114</td>
<td>0.041%</td>
<td>10295</td>
<td>0.004%</td>
<td>28427</td>
<td>0.011%</td>
</tr>
<tr>
<td>C5</td>
<td>285873232</td>
<td>143302</td>
<td>0.050%</td>
<td>13520</td>
<td>0.005%</td>
<td>33428</td>
<td>0.012%</td>
</tr>
<tr>
<td>C6</td>
<td>250477617</td>
<td>124598</td>
<td>0.050%</td>
<td>12487</td>
<td>0.005%</td>
<td>29668</td>
<td>0.012%</td>
</tr>
<tr>
<td>C7</td>
<td>262943930</td>
<td>114183</td>
<td>0.043%</td>
<td>9449</td>
<td>0.004%</td>
<td>29729</td>
<td>0.011%</td>
</tr>
<tr>
<td>M1</td>
<td>274060925</td>
<td>102049</td>
<td>0.037%</td>
<td>9693</td>
<td>0.004%</td>
<td>27210</td>
<td>0.010%</td>
</tr>
<tr>
<td>M2</td>
<td>278278868</td>
<td>100498</td>
<td>0.036%</td>
<td>9767</td>
<td>0.004%</td>
<td>28041</td>
<td>0.010%</td>
</tr>
<tr>
<td>M3</td>
<td>244772969</td>
<td>92624</td>
<td>0.038%</td>
<td>8963</td>
<td>0.004%</td>
<td>24944</td>
<td>0.010%</td>
</tr>
<tr>
<td>M4</td>
<td>206129129</td>
<td>69535</td>
<td>0.034%</td>
<td>6474</td>
<td>0.003%</td>
<td>20777</td>
<td>0.010%</td>
</tr>
<tr>
<td>M5</td>
<td>225964704</td>
<td>74778</td>
<td>0.033%</td>
<td>6892</td>
<td>0.003%</td>
<td>22713</td>
<td>0.010%</td>
</tr>
<tr>
<td>M6</td>
<td>215320045</td>
<td>73062</td>
<td>0.034%</td>
<td>6800</td>
<td>0.003%</td>
<td>22074</td>
<td>0.010%</td>
</tr>
<tr>
<td>M7</td>
<td>160726636</td>
<td>71249</td>
<td>0.044%</td>
<td>6580</td>
<td>0.004%</td>
<td>15182</td>
<td>0.009%</td>
</tr>
<tr>
<td>S1</td>
<td>192776425</td>
<td>68080</td>
<td>0.035%</td>
<td>6481</td>
<td>0.003%</td>
<td>19262</td>
<td>0.010%</td>
</tr>
<tr>
<td>S2</td>
<td>143660127</td>
<td>57229</td>
<td>0.040%</td>
<td>5469</td>
<td>0.004%</td>
<td>13270</td>
<td>0.009%</td>
</tr>
<tr>
<td>S3</td>
<td>218743879</td>
<td>76368</td>
<td>0.035%</td>
<td>7210</td>
<td>0.003%</td>
<td>21421</td>
<td>0.010%</td>
</tr>
<tr>
<td>S4</td>
<td>249773480</td>
<td>78422</td>
<td>0.031%</td>
<td>7215</td>
<td>0.003%</td>
<td>23664</td>
<td>0.009%</td>
</tr>
<tr>
<td>S5</td>
<td>140239637</td>
<td>49962</td>
<td>0.036%</td>
<td>4598</td>
<td>0.003%</td>
<td>12180</td>
<td>0.009%</td>
</tr>
<tr>
<td>S6</td>
<td>254749228</td>
<td>82105</td>
<td>0.032%</td>
<td>7282</td>
<td>0.003%</td>
<td>24364</td>
<td>0.010%</td>
</tr>
<tr>
<td>S7</td>
<td>194863138</td>
<td>68094</td>
<td>0.035%</td>
<td>6504</td>
<td>0.003%</td>
<td>19184</td>
<td>0.010%</td>
</tr>
</tbody>
</table>

### 3.4 Discussion

Although single phylogenetic marker based species definition are limited compared to whole genome based species definition, it still is the most effective and practical approach for microbial community analyses. Here we compare a single copy protein coding gene (*rplB*) with the commonly used SSU rRNA gene and demonstrate that single copy protein coding gene are suited for community species diversity analyses using all bacterial genomes and large soil shotgun metagenomes.
Figure 3.7: **Comparison of OTU numbers using SSU rRNA gene and rplB in large soil metagenomes.** The lower panel shows that there are more OTUs using SSU rRNA gene than rplB without even sampling, but the upper panel shows that there are less OTUs using SSU rRNA gene after even subsampling. X axis is percentage distance cutoff for OTU clustering and Y axis is OTU number.

Figure 3.8: **Comparison of SSU rRNA gene and rplB in beta diversity analysis (ordination) using large soil metagenomes.** Both genes show microbial community in corn rhizosphere is significantly different from those in Miscanthus and switchgrass. Additionally, rplB (both nucleotide and protein) also separates microbial communities of Miscanthus and switchgrass, suggesting the higher resolution of rplB.
Figure 3.9: Comparison of SSU rRNA gene and rplB in alpha diversity analysis (Shannon diversity index) using large soil metagenomes. Both genes show microbial community in corn rhizosphere is significantly less diverse than those in Miscanthus and switchgrass.

3.4.1 Choosing genomic datasets

For the comparison of rplB and the SSU rRNA gene within genomes, genome (assembly) quality is critical because erroneous sequences of SSU rRNA genes and rplB extracted from low quality assembly genomes could lead to problematic results. On the other hand, it is important to include as many genomes as possible to make the conclusion more generalizable. We tested two sets of genomes. The first set contained only completed genomes (genomes that have been closed as one sequence) (4457), which are expected to have higher quality. The second set is larger with all the available NCBI RefSeq Bacterial genomes (53865), most of which are not completed (with multiple contigs). The fact that the second set has a higher fraction of genomes with multiple copies of the rplB genes (expected in single-copy) and many E. coli strains in the second dataset show lower than 80% similarity for the SSU rRNA genes between strains (Fig. 3.5), imply, on average, lower quality genome assemblies in the second set. Thus we pick the first set (completed genomes) for our analyses. Short read sequencing technology has made sequencing bacterial genomes cheaper and faster, while not providing a parallel decrease in the cost of finishing a genome, so most assembled genomes are not complete [106]. Especially for genomic regions with multiple copies such as SSU rRNA genes, assembly from short reads are prone to assembly error and long-read sequencing technology such as PacBio and Oxford NanoPore are promising way to solve the problem.
3.4.2 Advantages of *rplB* over SSU rRNA gene as a phylogenetic marker

First, the SSU rRNA gene (a multiple copy gene) posts difficulties for interpreting species (OTU) abundance, while *rplB* (a single copy gene) does not have the same issue. Additionally, variations among multiple SSU copies can cause multiple OTUs (sequence clusters) from the same species [115], and thus lead to overestimation of species richness. Since, a single copy of the *rplB* gene is contained in every cell in a community, *rplB* the relative abundance of *rplB* gene sequences provides a mechanism for estimating the fraction of organisms possessing other less universal single-copy genes. Second, SSU rRNA genes are also more prone to assembly errors (chimera) than single copy genes due to their higher overall similarity and the presence of highly conserved regions interspersed in SSU rRNA genes. Multiple copies and variation among copies, which is also consistent with the results that there are unexpectedly low similarities (< 80%) observed among SSU rRNA genes of *E. coli* strains while *rplB* has high similarities (> 95%) among *E. coli* strains when incomplete genomes are included (Fig. 3.6). Note that these erroneous sequences might be further collected by databases and used as references for taxonomy, alignment, and chimera detection, and thus have an impact on common microbial ecology diversity analyses; switching to a single copy gene that is less prone to assembly error can mitigate the above problem. Third, we have shown the *rplB* gene is better able to differentiate closely related species than is the SSU rRNA gene, due to lower sequence similarity of the *rplB* gene compared to the SSU rRNA gene in pairwise comparisons of genome sequences (Figs. 3.2 to 3.5) and also *rplB* genes assembled from soil metagenomic data clustered into more OTUs than SSU rRNA genes indicating again more fine-grained resolution (Fig. 3.7). This is consistent with the crucial role SSU rRNA plays in translation as part of the ribosome (ensuring translation accuracy by working in concert with two other rRNA species, tRNAs, ribosomal proteins, translation initiation, elongation and termination factors etc. [111], which has also been confirmed by another study showing SSU rRNA genes (along with LSU rRNA genes, tRNA and ABC transporter genes) have been reported to be the most conserved genes [116]. Further, there are multiple cases of clearly different species with identical SSU rRNA gene (Fig. 3.2), and the heterogeneity among copies of SSU rRNA genes in
the same genome can be higher than those among different species [115], which means the SSU rRNA gene is not suitable to resolve lower taxonomies such as species and strains. And last, rplB has the advantage of having sequences in both nucleotide space and protein space. Higher level taxonomies can be resolved at protein space, and lower level taxonomies can be resolved at nucleotide space with protein sequences guiding the nucleotide alignment (codon alignment), which is critical for phylogenetic analysis and alignment-based OTU clustering.

3.4.3 Novelties in this study

Although there are other studies on comparing single copy housekeeping genes and SSU rRNA genes [40,41], the novelty in our analyses is that we use OTUs (from *de novo* clustering), which is more common than taxonomy in microbial diversity analyses and enabled by recent advancement in shotgun metagenomic analysis methodology, SSUsearch and EMIRGE (for SSU rRNA genes) [6, 114] and Xander (for rplB and other protein coding genes) [36]. In contrast, other studies do comparison at taxonomy level, which because existing reference sets cover only a small fraction of diversity, many reads cannot be classified at high confidence and are thus be dicarded, especially for genes other than SSU rRNA genes. Second, we compare these two genes using much larger genomic datasets and larger shotgun metagenomic datasets than were previously availabl. For example, Case et al. compared the rpoB gene (coding for the beta subunit of RNA polymerase) to the SSU rRNA gene using only 111 completed genomes from NCBI [40], while we include 4457 completed genomes, so can result in more generalizable conclusions. Similarly with shotgun metagenomic data, Roux et al. make comparisons using only a subset of samples from the Global Ocean Survey, which is about 6 billion bp in total from Sanger sequencing [41,117], while our soil metagenomic datasets are about 900 billion bp in total.

3.4.4 Lower strain level resolution of SSU rRNA gene due to its intra-genome variations

Additionally, we find inconsistent results when comparing two genes at strain level with different species (Figs. 3.3 to 3.5). The exception in *E. coli* could be explained by the intra-genome
variations of SSU rRNA genes (Fig. 3.1) that lead to larger inter-strain variation than \textit{rplB} among strains, which could not be interpreted as better resolution. Consistently, we observed OTU numbers from SSU decrease slower when distance cutoff is less than 4\% (slope of line in Fig. 3.7), which causes OTU number of SSU rRNA genes to be higher than \textit{rplB} protein and suggests that intra-genome variation of SSU rRNA genes are causing inflation of OTU numbers [115] and also make SSU rRNA genes not suitable for strain identification.

### 3.4.5 Comparison of two genes in \textit{de novo} OTU based community diversity analyses

Another goal of this study is to evaluate the effect of two different genes on common OTU based diversity analyses such as alpha diversity and beta diversity. In addition to more OTUs resulting from \textit{rplB} nucleotide sequences than SSU rRNA gene sequences using the same distance cutoff (0.03), diversity analyses of two genes provide the same biological conclusion, except that \textit{rplB} also separates \textit{Miscanthus} and switchgrass while it is not for corn, suggesting \textit{rplB} has higher resolution than the SSU rRNA gene. This is similar to comparison of taxonomy based diversity analysis and \textit{de novo} OTU based diversity analyses. Although species or genus level is enough to differentiate the biological treatments sometimes, higher resolution levels (OTU) are still preferred when the differences are not obvious, e.g., there is no difference in community composition at species level but strain compositions in some species are different. For the same reason, higher variation in \textit{rplB} makes the gene a better phylogenetic marker than SSU rRNA genes for analyzing more similar communities.

### 3.4.6 Sequencing depth needed for \textit{rplB} in metagenomes

Now that \textit{de novo} OTU based diversity analyses with \textit{rplB} in shotgun metagenomes have been enabled by Xander, another common question in experiment design is: How much sequencing is needed to ensure there is enough coverage on the \textit{rplB} gene for diversity analyses? Because of the high microbial diversity in soil and sequencing error, the thorough sequencing of soil metagenomes is difficult. Additionally, saturation of OTUs (species) in rarefaction curve is also not necessary
for beta-diversity analysis [54], so an empirical fold coverage of 3000 is recommended for soil samples based on experiences with the SSU rRNA gene [6], which means about 25 Gbp (3000*850 bp/0.01%) shotgun sequences are needed with 850 bp as the average length of rplB and 0.01% as the average percentage of rplB reads in total reads (Table 3.3).

### 3.4.7 Lack of universal primers for amplification and reference sequences

There are a few drawbacks to using single copy housekeeping genes such as rplB for diversity analysis. First, unlike SSU rRNA genes, which have highly conserved regions interspersed with highly variable regions, the redundancy in the genetic code, especially the high variability at the third codon position makes it difficult to design universal primers. Although better bulk assembly methods and new gene-targeted assembly methods (such as Xander) enable us to avoid primer amplification [6,36,114], the cost of shotgun sequencing of large numbers of samples at sufficient sequencing depth, as mentioned above, is still high and thus a limitation. Therefore, availability of universal primers are still desirable traits for a good phylogenetic marker. Second, rplB and other single copy housekeeping genes lack deep sequence reference sets, while the SSU rRNA gene has accumulated a large amount of reference sequences and dedicated databases, such as those in RDP and SILVA [23, 24] that are critical for diversity analyses.

### 3.5 Conclusion

Although genome level comparisons have the highest resolution, the phylogenetic marker method is still essential for community diversity analysis. We demonstrate that rplB, a single copy protein coding gene, can provide finer resolution to identify low level taxonomy such as species and strains than the more commonly used SSU rRNA gene (Fig. 3.2) and also finer scale de novo (OTU) diversity analysis (Figs. 3.7 and 3.8). In spite of the lack of references, more references are becoming available as bacterial genome sequencing becomes faster and cheaper [106]. In addition, as the switch to higher-quality long read sequencing for bacterial genome sequencing continues, the quality of these reference genomes will increase. Thus, single copy protein coding
genes such as *rplB* have great potential to complement or even replace the SSU rRNA gene as a phylogenetic marker.
CHAPTER 4

RHIZOSPHERE METAGENOMICS OF THREE BIOFUEL CROPS

4.1 Abstract

Soil microbes form beneficial association with crops in rhizosphere and also play a major role in ecosystem functions, such as N and C cycle. Crop roots had strong influences on the soil microbial community. Thus large-scale plantation of biofuel crops will have significant impact on ecosystem functions regionally and beyond. We compared rhizosphere microbial communities of corn (annual) and switchgrass and Miscanthus (perennials). This is the first comparative study of these biofuel crops using shotgun metagenomics and one of the largest sequencing efforts to date (about 1 TB bp in total). We compared the rhizosphere metagenomes at three levels: overall community structure (SSU rRNA gene), overall function (annotation from global assembly), and N cycle genes (from Xander). All three levels showed corn had a significantly different community from Miscanthus and switchgrass (except for AOA). In terms of life history strategy, the corn rhizosphere was enriched with more copiotrophs while the perennials were enriched with oligotrophs, which is further supported by higher abundance of genes in “Carbohydrates” and higher fungi/bacteria ratios. In addition, corn also had a less rich and even community, so the perennials managed to maintain a more diverse community even though investing less C in the rhizosphere. Moreover, a larger dispersion of corn data in ordination plots and enriched Penicillium (non-beneficial fungi) also indicate corn may not be doing as well in controlling its community and selecting beneficial member. Furthermore, the nitrogen fixing community of corn was dominated by Rhizobium (perhaps a legacy from prior legume crops) while the perennials had NifH sequences most related to Coraliomargarita, Novosphingobium and Azospirillum, indicating that the perennials can better select beneficial members. Moreover, higher numbers of genes for nitrogen fixation and lower number of genes for nitrite reduction suggest better nitrogen sustainability of the perennials. Thus our study provides comprehensive evidence showing perennial bioenergy crops have advantages
over corn in higher microbial species and functional diversity and in selecting members with beneficial traits, consistent with a higher level of sustainability of perennial biofuel crops.

4.2 Introduction

Bioenergy crops have the potential to produce renewable transportation fuels and with a net negative greenhouse gas production if managed appropriately [118], and to provide other ecosystem services [119, 120]. Bioenergy crop growth still requires some energy inputs such as N fertilizer, a component that the cropping system’s microbiome can potentially minimize. The microbiome, and especially the rhizosphere, is an important part of the soil where plants provide the microbes carbon and microbes provides the plant nutrients (N and P), growth hormones, and disease resistance [121–123]. Thus rhizosphere microbial community has the potential reduce the cost of growing bioenergy plants and is key for sustainability, especially when bioenergy crops are grown on marginal lands, i.e., not suitable for human food crops.

Vegetation type has an influence on microbial community structure through its root exudates and detritus [124, 125]. Here we studied the rhizosphere of three prominent biofuel crops (corn, Miscanthus and switchgrass) from the same family (true grasses) but varying markedly in phenotype. Miscanthus and switchgrass are perennial grasses with fibrous roots, while corn is an annual with less fibrous root. Miscanthus is the largest in size and is an exotic originating from tropical Asia, while switchgrass is native to US Midwest. Large scale planting of these biofuel crops may change soil microbial community structure and function, and impact global nutrient cycle processes. Corn typically needs fertilizer for maximum yields, and the N fertilizers are reported to cost about 40% of total energy input for the cropping system [126]. On the other hand, Miscanthus and switchgrass can grow with little or no N fertilizer [127–129], so the perennial grasses are thought to have better nitrogen sustainability. It is still not known how much rhizosphere microbes contribute to the better nitrogen sustainability of perennials since nutrient translocation before senescence and phyllosphere (leaf) microbes are also possible contributors. Here we not only study the overall microbial community structure of three biofuel crops but also focus on groups involved in the N
cycle. We hypothesize that the rhizosphere microbial communities of perennials are adapted to a low N fertilizer input environment and are enriched in N fixation genes, and has less denitrification genes.

Despite limited understanding of rhizosphere microorganisms, the chemistry of the nutrient cycles carried out by those microorganisms is relatively clear, including nitrogen fixation, nitrification, and denitrification. The nitrogen fixing microbial community is commonly measured with *nifH* as marker and the nitrifying community with AOA (archaeal *amoA*) and AOB (bacterial *amoA*) as their marker genes [47, 130]. Denitrification has multiple steps including nitrate reduction (Nar), nitrite reduction (Nir), nitric oxide reduction (Nor) and nitrous oxide reduction (Nos). Nir has two types, Cu-Nir (copper-based) and cd1-Nir (heme-based), encoded by *nirK* and *nirS*, respectively. Nor is encoded by *cNor* (with cytochrome c) or *gNor* (with quinone instead of cytochrome c). Nos also has different types, encoded by typical *nosZ* (clade I) and atypical *nosZ* (Clade II) [48, 49, 131, 132]. Thus amplicon based studies only targeting one single type of a gene for nitrification and denitrification likely underestimate abundance of the microbes involved.

Most surveys on N cycle related communities in soils use amplicon based methods [47,48,133], which has primer bias problems, especially for those N cycle genes that lack well curated reference databases [47, 48, 133]. In contrast, the shotgun metagenomic method produces short reads fragments representing all the genetic information of community members and thus has no primer bias. The challenge of the shotgun approach is the analysis of big data and its short read character [9,10]. There is a cross-disciplinary study showing the perennial grass can increase biodiversity across multiple taxa and enhance multiple ecosystems, but only methanotrophic bacteria is measured within microbiome [120]. Our study covers overall microbial diversity and also subgroups involved in N cycle that are of economic and environmental interest, and thus provides a more complete view of the microbiome. Further, we expect to see stronger microbial response by using rhizosphere soil (< 1 mm to root) rather than bulk soil used in other studies [120, 128].

In this study, we applied shotgun metagenomics (about 1 TB DNA data) in combination with ecology theory (life history strategy) to study rhizosphere microbial species and functional re-
response to three different crops (corn, *Miscanthus*, and switchgrass). We use computationally efficient methods to assemble globally (khmer) [10, 134, 135] and also extract genes of interest (SSUsearch and Xander) [6, 36] in this big data. Comparative metagenomics is applied at whole community species and function level and also N cycle gene level to understand microbial responses to different bioenergy crops.

4.3 Methods

4.3.1 Soil samples, DNA extraction, and sequencing

Rhizosphere samples of corn, switchgrass and *Miscanthus* were collected at Great Lake Bioenergy Research Center (GLBRC) intensive Cropping System Comparison site in Kellog Biological Station (KBS) (42.394828, -42.394828) ([http://data.sustainability.glbrc.org/pages/1.html](http://data.sustainability.glbrc.org/pages/1.html)) in Michigan on 10/12/2012. We sampled rhizosphere of each crop at seven plot areas. Each replicate was a composite of two or three plant close to the sampling spot. Roots were shaken vigorously to remove loosely attached soil, then cut off from stem and placed in plastic bags 4 degrees C. Soil closely attached to roots (< 1 mm) were collected and washed off roots as rhizosphere soil in lab (Fig. 4.1). Soil left were send to measure soil chemistry. DNA was extracted with PowerSoil DNA Isolation Kit from MO BIO, USA as described in [70] and then sent to Joint Genome Institute (JGI) for Illumina HiSeq 2500 sequencing (250 bp insert libraries and 2 x 150 bp reads).

4.3.2 Data preprocessing

Raw Illumina reads were quality trimmed using fastq-mcf (version 1.04.662, [http://code.google.com/p/ea-utils](http://code.google.com/p/ea-utils)) with flags “-l 50 -q 30 -w 4 -x 10 -max-ns 0 -X”. Then paired ends were merged by FLASH (version 1.2.7) [71] using flags “-m 10 -M 120 -x 0.20 -r 140 -f 250 -s 25” [6].
Figure 4.1: **Roots (Miscanthus) and rhizosphere soil.** Rhizosphere soil is collected by washing soil from the roots. The image shows that soil is very close (< 1 mm) to small roots.
4.3.3 Diversity analysis using SSU rRNA gene

SSU rRNA gene fragments in shotgun metagenome data were identified and an OTU table was made using fragments aligned to a 150 bp region of V4 (Escherichia coli position 577 to 727) by the SSUsearch pipeline [6] following a tutorial in https://github.com/dib-lab/SSUsearch. Ordination, richness and diversity analyses were calculated by phyloseq [136], simba [137] and vegan [138] packages in R [139].

4.3.4 Global metagenome assembly, annotation and functional diversity

Preprocessed reads of each crop were pooled and processed with digital normalization (-k 20 -C 10 -N 4 -x 2.14e11) and partitioning (-k 20 -N 4 -x 6e11) pipeline [7] following tutorials in http://nbviewer.jupyter.org/github/ngs-docs/ngs-notebooks/blob/master/ngs-70-hmp-diginorm.ipynb and http://nbviewer.jupyter.org/github/ngs-docs/ngs-notebooks/blob/master/ngs-71-hmp-partition.ipynb using khmer (version 0.6.1). The partitioned data were then assembled with velvet (version 1.2.08) [11] and SGA (version 0.9.35) [12]. For velvet, kmer size of 29, 39, 49, 59 and 69 were used and final assembly was merged with SGA. For SGA, overlap size of 29, 39, 49, 59, and 69 were used and final assemblies were clustered using cd-hit-est (-c 0.99) in CD-HIT [140]. Contigs shorter than 300 bp were discarded. Since velvet and SGA produced very similar assembly in terms of total size and longest contigs, the assembly from SGA was chosen for further analysis.

The assembly was then uploaded to MG-RAST [65] for gene calling and annotation. Gene calling files (with suffix .genecalling.coding.faa), clustering files (with suffix .cluster.aa90.mapping), and BLAT [84] tabular output files (with suffix .superblat.sims) were downloaded. BLAT hits with alignment longer than 30 amino acids, identity higher than 70%, and E-value lower than 0.00001 were used for downstream analysis. Number of hits to each reference in MG-RAST M5NR database for each sample was summarized based on BLAT hits, gene calling IDs, and clustering info. Then M5NR IDs were converted into subsystem IDs and further subsystem ontologies using MG-RAST API (http://api.metagenomics.anl.gov/api.html). A count table of sub-
system ontologies across each sample was made. Level 3 and Level 1 subsystems were used for ordination analysis with R phyloseq package [136]. For enrichment analysis, group_significance.py in QIIME [35] was used to find level 3 subsystems (pathways) that were significantly different between two crops (p < 0.05, n = 7) and those pathways with more than 10 total counts in all samples and a fold change less than 2/3 or larger than 3/2 were treated as significant.

4.3.5 N cycle gene assembly and diversity

Preprocessed reads of the same crop were combined and used as input to Xander (gene targeted metagenomic assembler) [36] with MAX_JVM_HEAP=300G, FILTER_SIZE=40, K_SIZE=45, genes=“nifH amoA_AOA amoA_AOB nirK nirS norB_cNor norB_qNor nosZ nosZ_a2 rplB” following tutorial in https://github.com/rdpstaff/Xander_assembler. Output files with taxon abundance (“*_taxonabund.txt”) were used for taxonomy-based analysis. The abundance of each N cycle gene was normalized by the number of rplB genes (single copy gene, per 10,000). For OTU-based analysis, assembled gene sequences were clustered by McClust [23]. Mean kmer coverage of each assembled sequence in each sample was calculated by KmerFilter.jar in Xander and then was used to adjust the abundance of OTUs in the sample. Diversity analysis was done by phyloseq and vegan package in R. Atypical nosZ (clade II) were divided into two groups (nosZ_a1 and nosZ_a2) due to large sequence variation within the group. However, nosZ_a1 were not yet included in Xander.

4.3.6 Data accession

The scripts used to analyze the data are available at https://github.com/jiarong/2015-glbrc/tree/master/scripts. Raw reads and assembled contigs can be accessed on JGI portal and MG-RAST respectively (Table 4.1).
Table 4.1: Size and JGI project ID of raw read data for 21 samples.

<table>
<thead>
<tr>
<th>Sample</th>
<th>JGI project ID</th>
<th>Data size</th>
</tr>
</thead>
<tbody>
<tr>
<td>C1</td>
<td>1023764</td>
<td>46</td>
</tr>
<tr>
<td>C2</td>
<td>1023767</td>
<td>39</td>
</tr>
<tr>
<td>C3</td>
<td>1023770</td>
<td>57</td>
</tr>
<tr>
<td>C4</td>
<td>1023773</td>
<td>53</td>
</tr>
<tr>
<td>C5</td>
<td>1023776</td>
<td>57</td>
</tr>
<tr>
<td>C6</td>
<td>1023779</td>
<td>51</td>
</tr>
<tr>
<td>C7</td>
<td>1023782</td>
<td>46</td>
</tr>
<tr>
<td>M1</td>
<td>1023785</td>
<td>56</td>
</tr>
<tr>
<td>M2</td>
<td>1023788</td>
<td>57</td>
</tr>
<tr>
<td>M3</td>
<td>1023791</td>
<td>50</td>
</tr>
<tr>
<td>M4</td>
<td>1023794</td>
<td>42</td>
</tr>
<tr>
<td>M5</td>
<td>1023797</td>
<td>45</td>
</tr>
<tr>
<td>M6</td>
<td>1023800</td>
<td>43</td>
</tr>
<tr>
<td>M7</td>
<td>1018623, 1018611</td>
<td>32</td>
</tr>
<tr>
<td>S1</td>
<td>1023803</td>
<td>40</td>
</tr>
<tr>
<td>S2</td>
<td>1018626, 1018614</td>
<td>28</td>
</tr>
<tr>
<td>S3</td>
<td>1023806</td>
<td>44</td>
</tr>
<tr>
<td>S4</td>
<td>1023809</td>
<td>49</td>
</tr>
<tr>
<td>S5</td>
<td>1018629, 1018617</td>
<td>27</td>
</tr>
<tr>
<td>S6</td>
<td>1023812</td>
<td>50</td>
</tr>
<tr>
<td>S7</td>
<td>1023815</td>
<td>39</td>
</tr>
</tbody>
</table>

4.4 Results

4.4.1 SSU rRNA gene based analysis

The average read number after quality trimming and paired end assembly was about 228 millions (range from 140 millions to 285 millions) for each replicate. The reads sizes ranged from 50 bp to 290 bp. An average of 0.039% of data were identified as SSU rRNA gene fragments (Table 4.2). The number of SSU rRNA gene fragments aligned to 150 bp of V4 region was 9.4% on average (range from 8.3% to 10.0%). All samples were subsampled to the same number (4598) of SSU rRNA gene fragments and a total of 12841 OTUs were clustered. The OTU number in each replicate was 2164 on average (range from 1707 to 2355).

Ordination analysis using OTUs clustered from fragments aligned to V4 region of SSU rRNA
gene showed that perennials (*Miscanthus* and switchgrass) had significantly different community structure from the annual (corn) (ADONIS: p < 0.001, $R^2 = 44.4\%$) (Fig. 4.2 A). Corn also had greater dispersion among replicates than the perennials (Fig. 4.2 B). Consistently, relative abundance of shared OTU and genera were highest between *Miscanthus* and switchgrass (Table 4.3). The three crops shared 21.5% of their OTUs, but 73.1% when abundance of each OTU was considered. At genus level, they shared 55.2% of genera, but 98.1% when abundance of each genus was considered.

Microbial communities in the rhizosphere of the perennials had higher species richness and diversity than did corn as measured by both alpha diversity (individual sample) and gamma diversity (samples merged by plant) (Figs. 4.2 and 4.17). Community members in perennials were also more evenly distributed than in corn (Fig. 4.3).

Corn had more Proteobacteria in rhizosphere while perennials had more Acidobacteria (Fig. 4.4).
Table 4.3: **Shared OTUs, genera, and contig among rhizosphere microbial communities of corn (C), Miscanthus (M), and switchgrass (S).** Miscanthus and switchgrass have more similarity to each other than to corn in OTU, genus, and contig.

<table>
<thead>
<tr>
<th></th>
<th>OTU</th>
<th>Genus</th>
<th>Contig</th>
</tr>
</thead>
<tbody>
<tr>
<td>C and M</td>
<td>0.34</td>
<td>0.8</td>
<td>0.17</td>
</tr>
<tr>
<td>C and S</td>
<td>0.33</td>
<td>0.8</td>
<td>0.17</td>
</tr>
<tr>
<td>M and S</td>
<td>0.4</td>
<td>0.82</td>
<td>0.26</td>
</tr>
</tbody>
</table>

Figure 4.2: **Species (OTU) and functional diversity comparison among three crops.** These abbreviations for the three crops, C (corn), M (Miscanthus), and S (switchgrass), are used throughout the paper. Corn has a different community composition from Miscanthus and switchgrass in both OTU (A) and function (level 3 subsystems) (subplot C), and corn has lower diversity in both OTU (B) and function (D). The symbol at top each graph indicates significance of difference (“***” indicates p < 0.001; “**” indicates p < 0.01; “*” indicates p < 0.05).
Figure 4.3: **Rank abundance curve of microbial communities of corn (C), Miscanthus (M) and switchgrass (S) using OTU from SSU rRNA gene.** Miscanthus and switchgrass have more even microbial community in rhizosphere than corn.

Top three genera are *Sphingomonas* (7.5%, Proteobacteria), *Penicillium* (3.5%, fungi), *Burkholderia* (3.4%, Proteobacteria) in corn, Da023 (5.0%, Acidobacteria), *Akiw543* (4.1%, Actinobacteria), *Sphingomonas* (3.1%, Proteobacteria) in Miscanthus, and *Da023* (6%, Acidobacteria), *Sphingomonas* (4.0%, Proteobacteria), and *Akiw543* (3.2%, Actinobacteria). Bacteria, Archaea, and Eukaryota were 91.9%, 0.7%, and 7.2% respectively on average across all samples (Table 4.4). Corn (4.9%) had more fungi than Miscanthus (2.2%) and switchgrass (2.0%). However, relative abundance of AMF (arbuscular mycorrhizal fungi) was lower in corn (0.1%) than Miscanthus (0.3%) and switchgrass (0.4%).
Figure 4.4: **Relative abundance of top ten most abundant bacterial phyla in three crops.** Corn (C) is enriched in Proteobacteria \( (p < 0.01, n = 7) \), and Miscanthus (M) and switchgrass (S) are enriched in Acidobacteria \( (p < 0.001, n = 7) \). Average of Proteobacteria in corn, Miscanthus, and switchgrass are 0.387, 0.282, and 0.312 respectively. Average of Actinobacteria in corn, Miscanthus, and switchgrass are 0.050, 0.101, and 0.117 respectively. Average of Acidobacteria in corn, Miscanthus, and switchgrass are 0.050, 0.101, and 0.117 respectively.

Table 4.4: **Average percentage of Bacteria, Archaea, Eukaryota, and fungi in rhizosphere communities of corn (C), Miscanthus (M) and switchgrass (S).** “AMF/fungi” is percentage of AMF (arbuscular mycorrhizal fungi) of total fungi. Corn has more fungi in the whole community but has less percentage of AMF within fungi.

<table>
<thead>
<tr>
<th></th>
<th>Bac</th>
<th>Arc</th>
<th>Euk</th>
<th>Fungi</th>
<th>AMF/fungi</th>
</tr>
</thead>
<tbody>
<tr>
<td>C</td>
<td>91.90%</td>
<td>0.70%</td>
<td>7.20%</td>
<td>4.90%</td>
<td>0.10%</td>
</tr>
<tr>
<td>M</td>
<td>94.10%</td>
<td>1.20%</td>
<td>4.50%</td>
<td>2.20%</td>
<td>0.30%</td>
</tr>
<tr>
<td>S</td>
<td>95.00%</td>
<td>0.90%</td>
<td>4.00%</td>
<td>2.00%</td>
<td>0.40%</td>
</tr>
</tbody>
</table>
4.4.2 Global assembly

We applied the digital normalization and partitioning method [7] to assemble pooled data (300 Gb) for each plant. The average computation time was 1510 CPU hours (63 days in total, Diginorm: 71h, Filterbelow: 25h, Loadgraph: 1166h, Partgraph: 205h, Mergegraph: 8h, Annograph: 27h, Extrgraph: 8h). Memory usage varied in different steps. The peak memory used was 1 TB in digital normalization, which produced an average false positive rate of less than 0.01 in k-mer counting (false positive rates below 20% are acceptable for digital normalization) [134]. Our assembly method produced assembly sizes ranging from 5.3 to 6.4 Gb with minimum contig size cutoff of 300 bp, and read mapping back rates (suggesting percentage of trimmed read data assembled) were similar among the three crops (Table 4.5). We also compared three crops based on sequence similarity among assemblies and found corn was the most different among three biofuel crops (Table 4.3).

Table 4.5: Global assembly statistics. Mapping% is the percentage of reads mapped to contigs. C stands for corn, M stands for Miscanthus, and S stands for switchgrass.

<table>
<thead>
<tr>
<th>Contig no. (M)</th>
<th>Total (Gbp)</th>
<th>Max (Kbp)</th>
<th>Median (bp)</th>
<th>Mapping%</th>
</tr>
</thead>
<tbody>
<tr>
<td>C</td>
<td>12.9</td>
<td>6.4</td>
<td>21</td>
<td>400</td>
</tr>
<tr>
<td>M</td>
<td>12.9</td>
<td>6</td>
<td>15.7</td>
<td>391</td>
</tr>
<tr>
<td>S</td>
<td>11.4</td>
<td>5.3</td>
<td>16</td>
<td>391</td>
</tr>
</tbody>
</table>

4.4.3 Functional diversity using annotation from assembly

After the assembled sequences were annotated by MG-RAST [65], we used “Subsystem” annotation for further analysis. Ordination analysis with both subsystem level 1 and 3 showed corn was different from the perennials and no significant difference between Miscanthus and switchgrass (ADONIS: p < 0.01). The total variation explained by crops was 74.7% and 77.2% at level 1 and 3 subsystems respectively (Figs. 4.2 and 4.5). The perennials had significantly higher in functional diversity (Shannon) at both levels (p < 0.001, n = 7).

Since the above ordination analysis showed corn was the most different and perennials were
Figure 4.5: **Functional diversity comparison among three crops using subsystem level 1.** Corn has a different community functional composition from Miscanthus and switchgrass (level 1 subsystem) (A), and corn has lower diversity (B). C stands for corn, M stands for Miscanthus, and S stands for switchgrass. Symbols at top each subplot indicates significance of difference (“***” indicates p < 0.001; “**” indicates p < 0.01; “*” indicates p < 0.05).

similar, we picked Miscanthus as a representative of perennials and focused on our comparison between corn and Miscanthus. We focused on level 1 and level 3 subsystems for enrichment analysis since level 1 should give us an overview of functions and level 3 should show pathway level information. Using level 1 subsystems, we found the rhizosphere of corn was enriched in “Iron acquisition and metabolism”, “Membrane Transport”, “Cell Wall and Capsule”, “Phages, Prophages, Transposable elements, Plasmids”, “Motility and Chemotaxis”, and “Carbohydrate”, while Miscanthus was enriched in “Amino Acids and Derivatives”, “Cell Division and Cell Cycle”, “Nucleosides and Nucleotides”, “Protein Metabolism”, “RNA Metabolism”, “Secondary Metabolism”, “Potassium metabolism”, “Respiration”, “Photosynthesis”, “Regulation and Cell signaling”, and “Stress Response” (FDR adjusted p < 0.05, n = 7).

At level 3, we found corn had more pathways significantly enriched belonging to level 1 subsystems of “Stress Response”, “Carbohydrate”, and “Cell Wall/Capsule”, while Miscanthus had more pathways significantly enriched in “Amino Acids and Derivatives”, and “Respiration” (Fig. 4.6). Within “Stress response”, “Desiccation stress” and “Oxidative stress” were significantly higher in corn, while “Cold shock” and “Osmotic stress” were significantly higher in Miscanthus. Within “Carbohydrates”, “Sugar alcohols”, “Aminosugars”, and “Monosaccharides” were enriched in corn, while “CO2 fixation”, “Central carbohydrate metabolism”, and “Polysaccharides”
were enriched in *Miscanthus*. Within “Phages, Prophages, Transposable elements, Plasmids”, “Gene Transfer Agent (GTA)”, “Plasmid related functions”, and “Phages and prophages” were enriched in corn, while “Transposable elements” were enriched in *Miscanthus*. Within “Nitrogen Metabolism”, “Nitrilase” and “Cyanate hydrosis” were enriched in corn, while “Dissimilatory nitrate reductase”, and “Denitrification”, “Allantoin Utilization”, “Nitrogen Fixation” were enriched in *Miscanthus* (FDR adjusted p < 0.05, n = 7).

### 4.4.4 N cycle gene diversity

All the N cycle genes except AOA showed significantly different among the three crops (Fig. 4.7). In terms of richness and diversity, *nosZ*, *nosZa2*, *cNor*, and *qNor* showed significant difference among three crops (Figs. 4.8 and 4.9. For the above four genes, their quantities were significantly higher in *Miscanthus* and switchgrass than in corn (p < 0.05, n = 7).

### 4.4.5 N cycle gene abundance and taxonomy composition

We used Xander [36] to assemble the N cycle genes and for the organism taxonomy of their nearest references. Figure 4.12 shows N cycle genes abundances, including nitrogen fixation genes (*nifH*), nitrification genes (AOA and AOB), and denitrification genes (*nirK*, *nirS*, *nosZ*, *nosZa2*, *qNor*, and *cNor*) normalized by *rplB* abundance. We found *nifH*, AOB, *nirK*, *nosZa2*, and *qNor* were significantly different between corn and *Miscanthus*. Additionally, *nirK*, *nosZ* and *qNor* were significantly different between corn and switchgrass. AOA and *nosZ* were significantly different between *Miscanthus* and switchgrass (p < 0.05, n = 7).

When combining the genes coding enzymes with the same function, nitrite reductase genes (*nirK* and *nirS*) are significantly more abundant in corn (C vs. M: p = 0.003, C vs. S: p = 0.05, one way t test across the three crops: p < 0.01), but nitric oxide reductase genes (*qNor* and *cNor*) and nitrous oxide reductase genes (*nosZ* and *nosZa2*) are significantly more abundant in the perennials (“norB” C vs. M: p = 0.001, “norB” C vs. S: p = 0.0004, “norB” one way t test across the three crops: p < 0.01, “nosZall” C vs. M: p= 0.0003, “nosZall” C vs. S: p = 0.1, “nosZall” one way t test:
Figure 4.6: **Function comparison between corn and Miscanthus.** Corn had more pathways significantly enriched belonging to level 1 subsystems of “Stress Response”, “Carbohydrate”, and “Cell Wall/Capsule”, while Miscanthus had more pathways significantly enriched in “Amino Acids and Derivatives”, and “Respiration”. Significance of enrichment is done by Kruskal-Wallis test. Those level 3 pathways with FDR (False Discovery Rate) adjust p < 0.05, total count larger than 10 and ratio of corn over Miscanthus (in log2 space) larger than 1.5 or smaller than -1.5 are chosen as significant.
Figure 4.7: **Ordination analysis of each N cycle gene and rplB using OTU.** All genes except AOA show significant separation of samples by corn (C), Miscanthus (M) and switchgrass (S). OTUs of each gene are clustered at 5% distance cutoff using protein sequence.

p < 0.01 (Fig. 4.10). The ratio of nitrification/fixation (amoA/nifH) and denitrification/nitrification ((nirK+nirS)/amoA, (qNor+cNor)/amoA, and (nosZ+nosZa2)/amoA) were not significantly different at \( \alpha = 0.05 \) level. However, the median of amoA/nifH was the highest in corn and the medians of (nirK+nirS)/amoA, (cNor+qNor)/amoA, and (nosZ+nosZa2)/amoA were highest in switchgrass (Fig. 4.11).

We also found a consistent pattern across all samples for those gene pairs coding enzymes with same function: AOA > AOB, nirK > nirS, nosZa2 > nosZ, qNor > cNor (Figs. 4.11 and 4.12). Moreover, nitrification gene (AOA and AOB) and denitrification gene (nirK and nirS, qNor and cNor, or nosZ and nosZa2) abundances were higher than nitrogen fixation genes (nifH), and denitrification gene abundances (nirK and nirS, qNor and cNor, or nosZ and nosZa2) were higher than
nitrification genes (AOA and AOB) (Figs. 4.11 and 4.12).

We then looked at detailed genus level distributions of *nifH*. There were 7, 9, and 5 genera in corn, *Miscanthus*, and switchgrass, respectively (Fig. 4.13). The taxonomic compositions of the communities were different among the three crops. On average, *Rhizobium*, *Bradyrhizobium*, and *Methylobacterium* were the most abundant in corn, *Coraliomargarita*, *Azospirillum*, and *Nostoc* were the most abundant in *Miscanthus*, and *Novosphingobium*, *Gluconacetobacter*, and *Azospirillum* were the most abundant in switchgrass (Fig. 4.13).

Since Xander assigned the taxonomy based on best hit to reference sequences, some assembled sequences might be quite different from its best-hit reference sequence and thus distant from its
Figure 4.9: Diversity (Shannon) comparison among corn (C), Miscanthus (M) and switchgrass (S) using OTU tables of N cycle genes. Symbols at top each subplot indicate significance of difference ("****" indicates p < 0.001; "***" indicates p < 0.01; "**" indicates p < 0.05, "." indicates p<0.1).

assigned taxon. Thus we examined the similarity distribution of our assembled sequences to references. Identities of assembled sequence to references range from 79.3% to 100% for nifH, 93.5% to 97.5% for AOA, 95.6% to 98.9% for AOB, 58.5% to 97.3% for nirK, 68.5% to 94.8% for nirS, 65.7% to 99.2% for nosZ, 54.7% to 97.7% for nosZa2, 71.1% to 98.4% for cNor, and 51.9% to 98.3% for qNor (Fig. 4.14). For nifH, the most abundant genera in corn, Miscanthus, and switchgrass has assembled sequences with lowest identity of 89.5% (Rhizobium), 83.1% (Coraliomargarita), 92.4% (Novosphingobium), respectively (Fig. 4.15).
Figure 4.10: **Denitrification gene abundances after combining the genes encoding enzymes with the same function.** “NirKS” = nirK and nirS. “NorB” = cNor and qNor. “NosZall” = nosZ and nosZa2. Each gene’s abundance is normalized relative to rplB copies (per 10,000). Three crops are significantly different in abundances of “nirKS”, “norB”, and “nosZall” (p < 0.05, n = 7). Corn is higher than perennials in abundance of “nirKS” but lower in abundance of “norB” and “nosZall”.

Figure 4.11: **Ratios of interest related to N cycle genes in corn (C), Miscanthus (M) and switchgrass (S).** “AmoA” is the sum of AOA and AOB; “nirKS” is the sum of nirK and nirS; “norB” is the sum of cNor and qNor; “nosZall” is the sum of nosZ and nosZa2. Each ratio is grouped by plant (7 replicates).
Figure 4.12: **Abundance and phylum level association of the indicated N cycle genes in corn (C), Miscanthus (M) and switchgrass (S).** Gene abundances are normalized relative to rplB copies (per 10,000). There are significant differences among three crops in nirK, nosZ, nosZa2 and qNor (p < 0.05, n = 7). Additionally, pairwise t test shows Miscanthus is significantly higher than corn in nifH and AOB (p < 0.05, n = 7).

### 4.4.6 Metadata

Soil chemistry analyses showed higher NO$_3^-$ and NH$_4^+$ in corn and no significant differences in SOM (Soil Organic Matter) and pH in soil from three (Table 4.6). Fold change of yield in 2012 over 2011 was significantly (p < 0.01, n = 5) lower in corn (more affected by drought) than perennials (Fig. 4.16 A). N$_2$O flux was higher in corn on average but not significant at $\alpha = 0.05$ level due to large variation in corn and negative values caused by low flux at sampling season (Fig. 4.16 B).

### 4.5 Discussion

In this study, we apply shotgun metagenomics to investigate rhizosphere microbial community structure and function of three biofuel crops - corn, Miscanthus, and switchgrass - with a focus on
Abundance and genus level distribution of \textit{nifH}. Y axis is gene abundance normalized relative to \textit{rplB} copies (per 10,000). X axis is sample labels in which first letter stands for crop (C for corn, M for \textit{Miscanthus}, and S for switchgrass). Three crops rhizospheres are quite different in genera with the closest sequence match. On average, \textit{Rhizobium}, \textit{Bradyrhizobium}, and \textit{Methylobacterium}-like sequences are the most abundant in corn, \textit{Coraliomargarita}, \textit{Azospirillum}, and Nostoc the most abundant in \textit{Miscanthus}, and \textit{Novosphingobium}, \textit{Gluconacetobacter}, and \textit{Azospirillum} the most abundant in switchgrass.

N cycle genes. We find corn is the most different among the three crops, in community structure (4.2 A), assembled genomic sequences (Table 4.3), and functional profile (4.2 C). It is expected that different plants have different rhizosphere microbial communities [124, 125], but our comparative metagenomic analyses of assemblies and functional profile provides information beyond the existing studies relying on 16S rRNA or a few other functional genes [125, 128]. Metagenome assembly could represent genomic content of communities and thus genetic diversity. Species diversity, genetic diversity, and functional diversity all show corn has significantly different rhizosphere microbial community than perennials (Fig. 4.2 A and C, Table 4.3), suggesting that large scale biofuel cropping can impact microbial communities and their associated ecosystem services.

The corn rhizosphere community also shows more variation among replicates in ordination plot
Figure 4.14: **Sequence identity of assembled N cycle genes to their best-hit references.** Percentage identities (Y axis) are grouped by plant (C for corn, M for Miscanthus, and S for switchgrass). The lower identity is, the more likely a assembled sequence is in correctly classified.

(Fig. 4.2 A). In contrast, community turnover within replicates (beta diversity) based on number of unique OTUs is similar among the three crops (Fig. 4.17), so the larger dispersion in corn is probably caused by variation in abundance of shared OTUs. There are two possible explanations: First, corn is annual and grows from seeds and establishes a new root system every year, while perennials (Miscanthus and switchgrass) have more stable root systems and probably a less varied community of varied rhizosphere community. Second, corn has been bred for grain yield under intensive management (e.g. fertilization) and some traits for recruiting a beneficial microbiome may have been lost. Thus larger variation of corn rhizosphere community may suggest weaker influence and selection from corn roots.

In spite of the larger dispersion in corn, perennials have higher alpha (local) diversity and
Figure 4.15: **Distribution of identity of assembled nifH sequences to their best-hit references.** Percent identities (Y axis) are grouped by genera. Assemblies with best-hit reference from Coraliomargarita and Frankia has percentage identity lower than other genera (< 85%), indicates these sequences are more likely to be incorrectly classified.

Figure 4.16: **Ratio of 2012 yield over 2011 and N₂O flux measured near sampling time.** Ratio of 2012 yield over 2011 of corn is significantly lower than Miscanthus and switchgrass, suggesting corn yield is impacted more by 2012 drought (A). Average of N₂O flux is higher in corn than Miscanthus and switchgrass, but not significant at $\alpha = 0.05$ level due to high variation in corn (B).
Table 4.6: Soil chemistry test results for 21 samples.

<table>
<thead>
<tr>
<th>Sample ID</th>
<th>NO3 ppm</th>
<th>NH4 ppm</th>
<th>% OM</th>
<th>Zn ppm</th>
<th>Mn ppm</th>
<th>Cu ppm</th>
<th>Fe ppm</th>
<th>pH</th>
</tr>
</thead>
<tbody>
<tr>
<td>C1</td>
<td>7.8</td>
<td>3.0</td>
<td>3.1</td>
<td>2.7</td>
<td>47.0</td>
<td>2.1</td>
<td>36.3</td>
<td>6.1</td>
</tr>
<tr>
<td>C2</td>
<td>1.7</td>
<td>3.4</td>
<td>2.8</td>
<td>1.9</td>
<td>36.0</td>
<td>1.5</td>
<td>28.7</td>
<td>5.7</td>
</tr>
<tr>
<td>C3</td>
<td>11.5</td>
<td>5.5</td>
<td>3.2</td>
<td>2.7</td>
<td>46.1</td>
<td>2.8</td>
<td>29.1</td>
<td>6.1</td>
</tr>
<tr>
<td>C4</td>
<td>4.7</td>
<td>3.0</td>
<td>3.3</td>
<td>2.9</td>
<td>42.5</td>
<td>2.6</td>
<td>22.3</td>
<td>6.2</td>
</tr>
<tr>
<td>C5</td>
<td>18.3</td>
<td>6.0</td>
<td>3.9</td>
<td>2.9</td>
<td>50.5</td>
<td>2.0</td>
<td>22.2</td>
<td>6.1</td>
</tr>
<tr>
<td>C6</td>
<td>4.6</td>
<td>5.5</td>
<td>3.9</td>
<td>3.1</td>
<td>38.8</td>
<td>2.6</td>
<td>27.4</td>
<td>6.2</td>
</tr>
<tr>
<td>C7</td>
<td>6.4</td>
<td>3.9</td>
<td>3.5</td>
<td>2.8</td>
<td>40.6</td>
<td>3.1</td>
<td>27.6</td>
<td>6.2</td>
</tr>
<tr>
<td>S1</td>
<td>1.2</td>
<td>3.6</td>
<td>3.6</td>
<td>2.3</td>
<td>32.0</td>
<td>2.1</td>
<td>21.5</td>
<td>6.2</td>
</tr>
<tr>
<td>S2</td>
<td>3.4</td>
<td>3.1</td>
<td>3.5</td>
<td>2.5</td>
<td>33.9</td>
<td>1.7</td>
<td>20.3</td>
<td>6.4</td>
</tr>
<tr>
<td>S3</td>
<td>2.7</td>
<td>3.1</td>
<td>4.0</td>
<td>2.9</td>
<td>34.2</td>
<td>2.7</td>
<td>25.0</td>
<td>6.2</td>
</tr>
<tr>
<td>S4</td>
<td>3.8</td>
<td>4.0</td>
<td>4.0</td>
<td>2.7</td>
<td>58.2</td>
<td>3.0</td>
<td>20.4</td>
<td>6.2</td>
</tr>
<tr>
<td>S5</td>
<td>1.1</td>
<td>2.9</td>
<td>4.9</td>
<td>3.1</td>
<td>56.3</td>
<td>3.1</td>
<td>22.9</td>
<td>6.1</td>
</tr>
<tr>
<td>S6</td>
<td>2.9</td>
<td>2.9</td>
<td>3.5</td>
<td>1.9</td>
<td>22.7</td>
<td>2.8</td>
<td>31.2</td>
<td>6.5</td>
</tr>
<tr>
<td>S7</td>
<td>1.1</td>
<td>3.1</td>
<td>3.6</td>
<td>2.2</td>
<td>27.0</td>
<td>2.3</td>
<td>28.1</td>
<td>6.1</td>
</tr>
<tr>
<td>M1</td>
<td>7.3</td>
<td>4.2</td>
<td>3.7</td>
<td>2.6</td>
<td>58.5</td>
<td>3.1</td>
<td>23.5</td>
<td>6.1</td>
</tr>
<tr>
<td>M2</td>
<td>2.4</td>
<td>4.9</td>
<td>3.5</td>
<td>1.8</td>
<td>34.5</td>
<td>3.6</td>
<td>25.0</td>
<td>5.8</td>
</tr>
<tr>
<td>M3</td>
<td>3.6</td>
<td>3.3</td>
<td>3.7</td>
<td>2.4</td>
<td>43.4</td>
<td>2.2</td>
<td>20.8</td>
<td>6.0</td>
</tr>
<tr>
<td>M4</td>
<td>1.5</td>
<td>3.2</td>
<td>3.9</td>
<td>3.1</td>
<td>48.6</td>
<td>3.1</td>
<td>25.5</td>
<td>6.0</td>
</tr>
<tr>
<td>M5</td>
<td>3.1</td>
<td>3.7</td>
<td>3.8</td>
<td>2.7</td>
<td>30.5</td>
<td>3.1</td>
<td>22.3</td>
<td>6.2</td>
</tr>
<tr>
<td>M6</td>
<td>2.1</td>
<td>4.8</td>
<td>4.3</td>
<td>3.5</td>
<td>51.9</td>
<td>3.3</td>
<td>23.8</td>
<td>6.2</td>
</tr>
<tr>
<td>M7</td>
<td>1.6</td>
<td>3.7</td>
<td>3.4</td>
<td>2.3</td>
<td>30.1</td>
<td>2.7</td>
<td>22.9</td>
<td>6.5</td>
</tr>
</tbody>
</table>

Also gamma diversity (when replicates are combined by plant) (Fig. 4.17). Thus change from annual to perennials as biofuel crops can increase the overall microbial diversity, which could also have a significant impact on the ecosystem. According to a study on disturbance and diversity models [141], evenness should increase when disturbance is introduced. Corn has to re-grow roots every year, which could be considered as a regular disturbance. The annual life cycle of corn should encourage a more even community, which contradicts to our result that corn is less even than perennials. Thus the lower evenness of corn rhizosphere community should be attributed to root exudates, root detritus (especially for corn), or agricultural management (e.g. fertilization), which are further discussed below.

Consistent with diversity, there are also significant differences in taxonomic composition between corn and the other two perennial crops. Proteobacteria, which was enriched in corn, are copi-
Figure 4.17: **Species (OTU) diversity of three crops.** *Miscanthus* and switchgrass are higher in both observed OTU number (q = 0 for “trudiv” function in simba package in R) (A) and another diversity index (B, q = 1 for “trudiv” function) (B). C stands for corn, M stands for *Miscanthus* and S stands for switchgrass. Alpha (diversity) refers to diversity of each replicate, gamma (diversity) refers to diversity when all replicates of each crop are combined, and beta (diversity) is ratio of gamma over alpha.

Oligotrophic taxa and Acidobacteria, which was enriched in perennials are oligotrophic taxa [142,143]. Increased N input can cause a shift from oligotrophic to copiotrophic taxa [144, 145]. Thus more copiotrophic taxa in corn are consistent with higher N fertilization and higher N found in its rhizosphere soil (Table 4.6). Further, since copiotrophs need more carbon as well nitrogen, we predict that corn is also providing more carbon source (root exudate and detritus) compared to perennials, which is consistent with more “Carbohydrates” related pathway genes enriched in corn rhizosphere as shown in the functional diversity analysis (Fig. 4.6). Since copiotrophs have a relatively faster growth rate in nutrient rich environment [144], they outnumbered the remainder, which explains the lower evenness in corn mentioned above (Fig. 4.3).

Additionally, the higher fungal-to-bacterial ratio in corn also suggests corn provides more carbon source to the rhizosphere compared to perennials and encourages copiotrophic bacteria, which is consistent with the above. Higher fungal-to-bacterial ratio commonly indicates higher C/N ratio since fungal biomass has higher C/N ratio than bacteria [94, 95]. Due to the fact that fungi has
larger biomass to DNA ratio than bacteria (some fungal hyphae may not have nuclei) and fungal DNA are harder to extract [146], the ratios (DNA based) here are much smaller than reported in other studies using biomass [70]. Moreover, *Penicillium*, a commonly saprotroph but hardly known as beneficial to plants (to the best of our knowledge), is highly enriched in corn (the third most abundant genus), which suggests that corn provides more carbon source to rhizosphere but not effectively recruiting beneficial members (selecting non-beneficial ones in this case).

There is more “Stress responses” related pathway genes enriched in corn, which indicates rhizosphere community of corn is under more stress. Especially, the enriched are “Desiccation stress” related genes which is consistent with 2012 being a drought year and the corn yield was affected by the drought the most (Fig. 4.16 A). Moreover, higher number of phage and prophage related genes in corn rhizosphere also indicates that corn rhizosphere community is at higher risk of phage infection and thus unstable, consistent with lower evenness and diversity of corn rhizosphere community compared to perennials.

Our global assembly and annotation pipeline is not suitable for evaluating N cycle genes, since the annotation database that we use, SEED subsystem [65], does not include any nitrification genes. Moreover, Xander provides better sensitivity and specificity, and recovers longer assemblies [36]. OTU based ordination analyses of all N cycle genes except AOA shows the significant separation of all three crops (Fig. 4.7), which is the same as SSU rRNA gene and functional profile (Fig. 4.2 A and C) and suggests that nitrogen fixing, nitrifying, and denitrifying community are also changed when crops are switched from corn to *Miscanthus* and switchgrass, consistent with the overall community, except that the separation of *Miscanthus* and switchgrass becomes more significant.

Perennials has more *nifH* than corn on average (significant for *Miscanthus* but not significant for switchgrass at $\alpha = 0.05$ level), suggesting perennials have more nitrogen fixing microbes, which could explain why *Miscanthus* and switchgrass growth does not respond to nitrogen fertilization [127, 129]. Our collaborators in KBS (where our samples were collected) find switchgrass has not responded to fertilization after 7 years (S. Roley, unpublished data). This is also consistent with another study using qPCR [128]. Three crops are very different in taxa with nearest se-
quence match with *Rhizobium*-like sequences as the most abundant in corn, *Coraliomargarita*-like sequences as the most abundant in *Miscanthus*, and *Novosphingobium*-like sequences as the most abundant in switchgrass (Fig. 4.13), which suggests each crop has selected different nitrogen fixing members, consistent with ordination analysis using OTUs (Fig. 4.7). Since these plots were planted with soybean and alfalfa before the experiment site was established, *Rhizobium* and *Bradyrhizobium*, also enriched in corn samples, were present before the biofuel crops were planted. On the other hand, perennials can selected N-fixing microbes such as *Coraliomargarita*, *Azospirillum*, *Nostoc*, *Novosphingobium*, and *Gluconacetobacter*, not Rhizobia. Additionally, note that there is a lot of variation in taxonomic composition in replicates (Fig. 4.13), but clearly there are different genera that stand out for all three crops though weak in a replicate or two, which underscores the importance of replication in soil studies.

Two nitrifying groups (AOA and AOB) both have very low diversity (three OTUs) compared to other genes (Figs. 4.8 and 4.9). The AOB communities were significantly different among the three crops, while AOA were not (Fig. 4.7), indicating that AOB community composition were more responsive to crop, consistent with a study using amplicon methods [147, 148]. Further, the result that corn with added N fertilizer does not have the highest abundance of nitrifying members (Fig. 4.12) suggests nitrifier abundance does not positively respond to N fertilizer, which may be explained by drought that caused too low soil moisture for nitrifiers. Conversely, the perennials were less affected by the drought due to their deeper and larger root systems.

Denitrification has multiple steps (nitrite reduction, nitric oxide reduction, and nitrous oxide reduction included our analysis) and each step has more than one gene coding enzymes with the same function, which makes it difficult to draw functional conclusions when comparing the three crops. For example, *nirK* is significantly higher in corn but *nirS* is not significantly different among the three crops. Thus we combined those genes coding enzymes with the same function, which leads to the finding that there are significantly more nitrite reductase genes but less nitric oxide reductase genes and nitrous oxide reductase genes in corn (Fig. 4.10). Since nitrite reduction is considered as the rate limiting step in denitrification [131] and the abundance of nitrite reduction genes are
significantly lower than downstream (Fig. 4.10), corn has selected a higher overall denitrification potential, which is consistent with more N fertilizer input in corn (higher substrate availability) and higher average N₂O flux in corn (Fig. 4.16 B) [149].

Other than community structure and function comparisons among the three crops, we also have some general findings about rhizosphere soil at our sampling site. First, ammonia-oxidizing archaea were about three times more abundant than ammonia-oxidizing bacteria on average, consistent with other studies on soil with similar conditions [130, 150, 151]. Bacteria with nirK type nitrite reductase are about nine times more abundant than those with nirS. Bacteria with qNor type nitric oxide reductase are 14 times more abundant than those with cNor. Bacteria with nosZa2 are four times more abundant than nosZ (Fig. 4.11). The above ratios also suggest that traditional amplicon studies that only target one gene in those processes might yield misleading conclusions, e.g. nosZ of nitrous reduction was higher in corn but nosZa2 was higher in Miscanthus and switchgrass. Second, nitrogen fixing bacteria were about 0.4% of total community on average; nitrifiers were about 1.4% on average; denitrifiers were about 10.6%, 22.4% and 15.6% of total community for nitrite reduction, nitric oxide reduction, and nitrous oxide reduction, respectively (Figs. 4.10 and 4.12). Both the above ratios and abundance are valuable information for future studies of this sampling site.

Another contribution of this study is that we find full length N cycle genes that are distant from existing references (Fig. 4.14), which can help improve primer design for these genes. Further, since taxon assignment is done by best hit reference sequence, taxonomy information of the assembled N cycle gene may be not correct (especially for those with low identity), but suggests these genes are in taxa whose genomes have not yet been sequenced (or perhaps the strains not yet isolated). It is also worth mentioning that the most abundant group of assembled nifH sequences (those assigned to the most abundant genus Coraliomargarita) in Miscanthus have an identity of only about 85% to best hit reference, which is from coral reefs (Fig. 4.15). This suggests that a major nitrogen-fixing strain in soil remain to be discovered.
4.6 Conclusion

In this study, we showcase the power of shotgun metagenomics to study microbial ecology at both community structure and function levels. We sequenced about 1 TB of rhizosphere metagenome, which is one of the largest sequencing efforts on rhizosphere soil and enables us to study functional traits that are not abundant. Overall community structure (SSU rRNA gene), overall function (annotation from global assembly), and N cycle genes (except AOA) all show the corn rhizosphere microbiome was significantly different from Miscanthus and switchgrass, suggesting changing bioenergy crop from corn (annual) to Miscanthus and switchgrass (perennial) will have impact on ecosystem functions carried out by microbiome. Further, we find corn have less influence and selection on its rhizosphere community, supported by larger variation in community composition, enriched Penicillium (non-beneficial fungi), and predominance of Rhizobium and Bradyrhizobium (left from before establishing the study site). Further, perennials manage to maintain more diverse rhizosphere microbial communities by investing less carbon source (root detritus and exudate). Moreover, we find perennials have more N fixing genes (nifH) and nitrite reducing genes (sum of nirK and nirS), which agrees with better N sustainability of Miscanthus and switchgrass [127, 129]. Thus perennials bioenergy crops have advantage over corn in maintain microbial species and functional diversity and also selecting members with beneficial traits.
BIBLIOGRAPHY
BIBLIOGRAPHY


