# Understanding lignocellulosic deconstruction using bioinformatics tools

By

JOSHUA THOMAS CLAYPOOL

M.S. (Iowa State University) 2013

B.S. (University of California, Davis) 2011

DISSERTATION

Submitted in partial satisfaction of the requirements for the degree of

DOCTOR of PHILOSOPHY

In

Biological Systems Engineering

in the

OFFICE OF GRADUATE STUDIES

of the

UNIVERSITY OF CALIFORNIA

DAVIS

Approved:

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

Christopher Simmons, Chair

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

Jean S VanderGheynst

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

Shrini Upadhyaya

Committee in Charge

2017

# Abstract

# Dedication

Contents

[Understanding lignocellulosic deconstruction using bioinformatics tools i](#_Toc474219171)

[Abstract ii](#_Toc474219172)

[Dedication iii](#_Toc474219173)

[List of Tables viii](#_Toc474219174)

[List of Figures ix](#_Toc474219175)

[Chapter 1 Introduction 1](#_Toc474219176)

[1.1 Background and Motivation 1](#_Toc474219177)

[1.2 16S rRNA gene sequencing 4](#_Toc474219178)

[1.3 Metagenomics 5](#_Toc474219179)

[1.4 Network analysis for studying microbial ecology 6](#_Toc474219180)

[Chapter 2 Characterization of bacterial communities for high-solids deconstruction of lignocellulosic tomato processing waste 9](#_Toc474219181)

[2.1 Abstract 9](#_Toc474219182)

[2.2 Introduction 11](#_Toc474219183)

[2.3 Materials and Methods 13](#_Toc474219184)

[2.3.1 Bioreactor Preparation and Operation 13](#_Toc474219185)

[2.3.2 Enzyme Extraction and Activity Assay 15](#_Toc474219186)

[2.3.3 DNA Extraction and Sequencing of the16S rRNA gene V4 Hypervariable Region 16](#_Toc474219187)

[2.3.4 Processing and Analysis of Sequencing Reads 16](#_Toc474219188)

[2.3.5 Data Analysis 17](#_Toc474219189)

[2.4 Results 18](#_Toc474219190)

[2.4.1 Microbial respiration and endoglucanase activity 18](#_Toc474219191)

[2.4.2 Taxonomy 20](#_Toc474219192)

[2.4.3 Metagenome 23](#_Toc474219193)

[2.5 Discussion 32](#_Toc474219194)

[2.6 Conclusion 39](#_Toc474219195)

[Chapter 3 Network meta-analysis method development and validation 40](#_Toc474219196)

[3.1 Abstract 40](#_Toc474219197)

[3.2 Introduction 41](#_Toc474219198)

[3.3 Materials and Methods 44](#_Toc474219199)

[3.3.1 Simulated Microbial Community Construction 44](#_Toc474219200)

[3.3.2 Network Construction 44](#_Toc474219201)

[3.3.3 Correlation Combination 45](#_Toc474219202)

[3.3.4 P-value Combination 45](#_Toc474219203)

[3.3.5 Comparison of Results 45](#_Toc474219204)

[3.4 Results 46](#_Toc474219205)

[3.4.1 Network recovery 46](#_Toc474219206)

[3.4.2 Sparsity 48](#_Toc474219207)

[3.5 Discussion 50](#_Toc474219208)

[3.6 Conclusion 52](#_Toc474219209)

[Chapter 4 Meta-analysis of solarization communities using network analysis 53](#_Toc474219210)

[4.1 Abstract 53](#_Toc474219211)

[4.2 Introduction 54](#_Toc474219212)

[4.3 Methods 56](#_Toc474219213)

[4.3.1 DNA-sequencing and Analysis 56](#_Toc474219214)

[4.3.2 Correlations 56](#_Toc474219215)

[4.3.3 Network Construction and Analysis 56](#_Toc474219216)

[4.3.4 Predicted Metagenomics 56](#_Toc474219217)

[4.3.5 Statistical Analysis 58](#_Toc474219218)

[4.4 Results 58](#_Toc474219219)

[4.4.1 Network Structure 58](#_Toc474219220)

[4.4.2 Network Characterization 59](#_Toc474219221)

[4.4.3 Solarization Results 61](#_Toc474219222)

[4.4.4 Metagenome 66](#_Toc474219223)

[4.5 Discussion 69](#_Toc474219224)

[4.6 Conclusion 70](#_Toc474219225)

[Chapter 5 70](#_Toc474219226)

[5.1 Abstract 70](#_Toc474219227)

[5.2 Introduction 70](#_Toc474219228)

[5.3 Methods 70](#_Toc474219229)

[5.4 Results 71](#_Toc474219230)

[5.5 Discussion 71](#_Toc474219231)

[5.6 Conclusion 71](#_Toc474219232)

[Chapter 6 Concluding Remarks 71](#_Toc474219233)

[APPENDIX 1 71](#_Toc474219234)

[APPENDIX 2 71](#_Toc474219235)

[APPENDIX 3 71](#_Toc474219236)

[References 72](#_Toc474219237)

# List of Tables

[Table 2.1: Cumulative CO2 respiration from mesophilic aerated, mesophilic non-aerated, thermophilic aerated, and thermophilic non-aerated. Different letters indicate significant differences between treatments. DW= dry weight 19](#_Toc477946009)

[Table 2.2: ANOVA table of sugars released during the hydrolysis of Carboxymethyl cellulose (CMC) from extracted enzymes 20](#_Toc477946010)

[Table 2.3: ANOVA table from the result of a permutation test on the Constrained Correspondence Analysis from the OTU’s present in each samples 22](#_Toc477946011)

[Table 2.4: Top five OTU relative abundances from the final time point of each community. Numbers indicate the Greengenes ID, single letter represents the phylum (F – Firmicutes, P – Proteobacteria, T – Tenericutes), and italicized taxonomy represents the highest degree of discernible classification obtainable using 97% similarity. 24](#_Toc477946012)

[Table 4.1: Primary author, tested substrate, number of samples utilized from experiment, and year the experiment was run 56](#_Toc477946013)

[Table 5.1: Sample metadata for construction of artificial neural network 82](#_Toc477946014)

[Table 5.2: Parameters and Equations for modeling of in-field microbiome using ANN 83](#_Toc477946015)

[Table 5.3: Number of samples in each error cluster identified by MClust 85](#_Toc477946016)

# List of Figures

[Figure 1.1: Approximate representation of classification accuracy using a sliding average with variable regions denoted above the gene position. See Wang et al. for original work [24]. 5](#_Toc477946017)

[Figure 1.2: All types of possible direction independent ecological interactions between hypothetical OTU’s A and B. The colors reinforce the sign for positive (green), negeative (red), and neutral (white/blue). 8](#_Toc477946018)

[Figure 1.3: (A) Networks are constructed based on connections between different OTU’s based on different correlation methods and the strength of the correlation is depicted by the thickness of the connection between the nodes (OTU’s). (B) Community detection methods will attempt to divide the network into smaller sub-communities as depicted by the different colors. 9](#_Toc477946019)

[Figure 2.1: Sugars released as measured by DNS assay from extracted enzymes on CMC. 20](#_Toc477946020)

[Figure 2.2: Plot of Constrained Correspondence Analysis. 21](#_Toc477946021)

[Figure 2.3: Relative abundances of the initial, T1, and T2 communities. “p\_” indicates we are referring to the Phyla. Phyla present in all communities at greater than 2% are shown whereas the remaining phyla are classified as “Other”. 23](#_Toc477946022)

[Figure 2.4: Principal Component Analysis of significant genes at T1 (A) or T2 (B). 26](#_Toc477946023)

[Figure 2.5: Relative abundance of target genes on target substrates for compost, T1, and T2. 27](#_Toc477946024)

[Figure 3.1: Figures A-E are organized by number of samples included in the analysis (rows) and method applied (columns). Informedness was plotted showing areas of where filtering produces a more/less informed network. Fisher’s Z transformation is shown after combining 10 total samples (A) and 20 total samples (B). Average R is shown for 10 (B) and 20 (E) samples respectively. All are compared to running all 20 samples through SparCC (C). 48](#_Toc477946025)

[Figure 3.2: Figures A-E are organized by number of samples included in the analysis (rows) and method applied (columns). Informedness was plotted showing areas of where filtering produces a more/less informed network. Fisher’s method for p-value combination is shown after combining 10 total samples (A) and 20 total samples (B). Average P is shown for 10 (B) and 20 (E) samples respectively. All are compared to running all 20 samples through SparCC (C). 49](#_Toc477946026)

[Figure 3.3: Figures A-F are organized by sparsity (rows) and method applied (columns). Informedness was plotted showing areas of where filtering produces a more/less informed network. Figures A, C, and E are output from SparCC without having been split prior. Figures B, D, and F have been split systematically to represent different environments and then were recombined using Fisher’s method for p-value combination and Fisher’s Z transformation 51](#_Toc477946027)

[Figure 4.1: Distribution of OTU's in solarization interactome (p-value ≤ 0.1; correlation ≥ 0.6) 61](#_Toc477946028)

[Figure 4.2: Solarization network (Gephi v 0.9.1) with modules detected by Louvain method shown as different colors 62](#_Toc477946029)

[Figure 4.3: Heatmap of clusters present by treatment and depth in the studies included in the construction of the interactome. 64](#_Toc477946030)

[Figure 4.4: Correlations of clusters to both individual and total volatile fatty acid production 65](#_Toc477946031)

[Figure 4.5: Zi-Pi plot displaying the connectedness of OTU's. Green circles (●) represent hubs, red circles (●) represent connectors, and blue circles (●) represent cornerstone OTU’s. 66](#_Toc477946032)

[Figure 4.6: Heatmap of the top 5 hubs (A) and top 10 connectors (B) by relative abundance in both depth and sample. 67](#_Toc477946033)

[Figure 4.7: Heatmap of relative abundance of genes within each cluster using 1000 counts per OTU to generate total counts for (A) deconstructive genes and (B) fermentative genes. 69](#_Toc477946034)

[Figure 4.8: Correlation of (A) deconstructive and (B) fermentative genes to VFA production 70](#_Toc477946035)

[Figure 5.1: Measured versus the prediction of ANN model 84](#_Toc477946036)

[Figure 5.2: Distribution of errors by cluster as shown by module 85](#_Toc477946037)

[Figure 5.3: Raw (A) and normalized (B) output from ANN predictions using equations in Table 5.2 The black line indicates current temperature associated with diurnal fluctuations as mimicked from Simmons et al. [133]. 86](#_Toc477946038)

# Introduction

## Background and Motivation

Microbial communities have become of increasing interest as the diversity of environments they are found continue to expand [[1](#_ENREF_1)]. This has often prompted the question as to what are the microorganisms doing to survive in these various environments. With the advent of next-generation sequencing (NGS), information regarding microbial communities, many currently uncultivable, has become more accessible than ever [[2](#_ENREF_2)]. The NGS has allowed us to identify organisms using16S rRNA gene sequencing, the genetic potential of the community using metagenomics, and what pathways organisms are utilizing by using metatranscriptomics and this has started to offer information regarding the operations and survival mechanisms in diverse environments [[3-5](#_ENREF_3)]. As our knowledge of microbial communities grows, both our ability is increasing as well as our interest to engineer these communities [[6](#_ENREF_6), [7](#_ENREF_7)]. A field that shares these interests is that of lignocellulosic deconstruction in high solids environments [[4](#_ENREF_4)]. Deconstructing lignocellulosic matter has many areas of application, ranging from biofuels, agricultural applications, all the way to global carbon cycling but currently suffers from slow rates and inhibitory by-products [[8-11](#_ENREF_8)].

Microbial communities naturally deconstruct lignocellulosic matter in these high solids environments in the form of plant litter, dead plant material that has fallen to the ground. Lignocellulosic matter however is a complex mixture of polymers typically divided up into cellulose, hemicellulose, lignin, and sometimes pectin. Cellulose represents an excellent substrate for most deconstructive microorganisms, but access to this energy-rich polymer is often hindered by lignin and hemicellulose. While most of the cell wall components can be readily degraded by bacteria, lignin is a biologically recalcitrant material thought to mostly be degraded by fungi although there are known bacterial pathways [[5](#_ENREF_5), [11-13](#_ENREF_11)]. Although lignin is difficult to degrade, many microorganisms possess a suite of enzymes, such as laccase or versatile peroxidase, designed to degrade lignin and the other plant cell wall components that provide access to the cellulose. Understanding what microorganisms are responsible for the degradation of organic matter starts to provide a new understanding to how the plant matter is broken apart.

Methods to understand how these microbial communities are changing often examine how a community shifts after the introduction of some selective pressure. Selective pressures can include such environmental factors as temperature, oxygen concentration, substrate, solids content, and pH [[10](#_ENREF_10), [14](#_ENREF_14), [15](#_ENREF_15)]. In the environment not all selective pressures are known, and it can be valuable to conduct lab trials to understand where shifts may have occurred. Often the pursuit of sequencing a metagenome, a genome of all the microorganisms, can provide valuable insight into community function by identifying genes and pathways relevant to survival in the relative environment [[4](#_ENREF_4), [16](#_ENREF_16)]. If possible performing a complete “omics” profile (genomics, transcriptomics, proteomics, secretomics) can provide one of the most comprehensive understanding of what is occurring in the environment, but can be resource intensive to obtain and difficult to analyze. Most microbial ecology studies will conduct 16S rRNA gene sequencing to identify relative abundance of community members. At the heart of all these studies is attempting to understand how the present organisms are working together to accomplish a specific goal. The ultimate goal being able to engineer a designed microbial community which would open new possibilities in industrial biotechnology or understand host-microbiome responses in a more controlled manner [[6](#_ENREF_6), [7](#_ENREF_7), [17](#_ENREF_17)].

Simple synthetic communities have been studied in the saccharification using both bacteria and fungi to deconstruct different straws with a predominant finding that the interactions had a positive effect on sugar release [[18](#_ENREF_18)]. This study did show that not all combinations resulted in an additive effect and that perhaps a more intricate set of interactions was at play and did not elucidate exact mechanisms for the apparent cooperation between two isolates. As we start to understand more about microbial communities, a major area of intense research is aimed at understanding microbial community interactions [[6](#_ENREF_6), [7](#_ENREF_7)]. To this effect, methods that researchers are using to pursue this information are construction of synthetic microbial consortia and network analysis. Synthetic microbial communities are proving to be their own challenge as many relevant species have not been previously cultured and/or cannot be cultured individually [[7](#_ENREF_7)]. When this occurs, network analysis is an excellent opportunity to research the microbial interactions and start to develop insight prior to pursuing cultivation techniques.

Network analysis for microbial ecology opens up new doors, just as in social network analysis, to identify sub-communities and unique organisms that have with unique properties [[19](#_ENREF_19), [20](#_ENREF_20)]. So rather than rely on isolation techniques to provide new information to how a community operates, network analysis attempts to identify organisms through construction of correlations between individual microorganisms using next generation sequencing (NGS) and correlation detection software. A problem associated with this technique relates primarily to sparsity, presence of zeroes, that can increase false correlations and the ensuing network become less interpretable [[21](#_ENREF_21)]. This introduces problems in the event of combining studies or even data from environmentally dissimilar conditions due to either tossing out of data or creating false connection due to environmental niches rather than a true interaction [[22](#_ENREF_22)]. Improving techniques that can circumvent sparsity in a typical microbial community dataset could prove useful in finding the true interactions present. Nevertheless, networks can provide new information on how microbial communities function that may lead to eventual construction or control of microbial communities.

## 16S rRNA gene sequencing

16S rRNA gene sequencing involves sequencing of a region in the microbial genome, approximately 1600 base pairs in length (Figure 1.1), that is hypervariable interspecies and hyperconserved intraspecies [[23](#_ENREF_23), [24](#_ENREF_24)]. Most of these studies aim to sequence only one or two of the nine hypervariable regions to provide the information for taxonomic studies although original apsects of microbial communities would sequence dominant organisms and use BLAST, a basic local alignment search tool [[25](#_ENREF_25)]. When the sequences are identified, typically the sequence is identified with some identifier label “OTU X”, where X is some number cataloged in a database, and OTU in this case refers to operational taxonomic unit, or best available taxonomic classification for this sequence. The number of available OTU’s in database are growing exponentially with decreasing costs of sequencing and can provide valuable insight into other occurrences associated with OTU’s identified in new studies. While these methods have elucidated much new information in the study of microbiomes, there are still many active issues that relate to sequencing processing, OTU binning, and taxonomic calling. Sequencing of bacterial ribosomal genes has proved to be effective for identifying many operational taxonomic units (OTU) in many different environments [[5](#_ENREF_5), [26](#_ENREF_26), [27](#_ENREF_27)]. As general sequencing guidelines for microbial communities, it is required that there is greater replication of unique sequences than length of sequence to establish differences in between communities at the sacrifice of improved identification of the OTU in question. To combat this problem, studies have examined the entire 16S rRNA gene for ability to classify OTU’s using different sections and lengths of the 16S rRNA gene (Figure 1.1) [[23](#_ENREF_23), [24](#_ENREF_24)] . The consensus amongst studies show the V4 region to be the shortest geodesic path as compared to sequencing the entire 16S rRNA gene [[23](#_ENREF_23), [24](#_ENREF_24)].

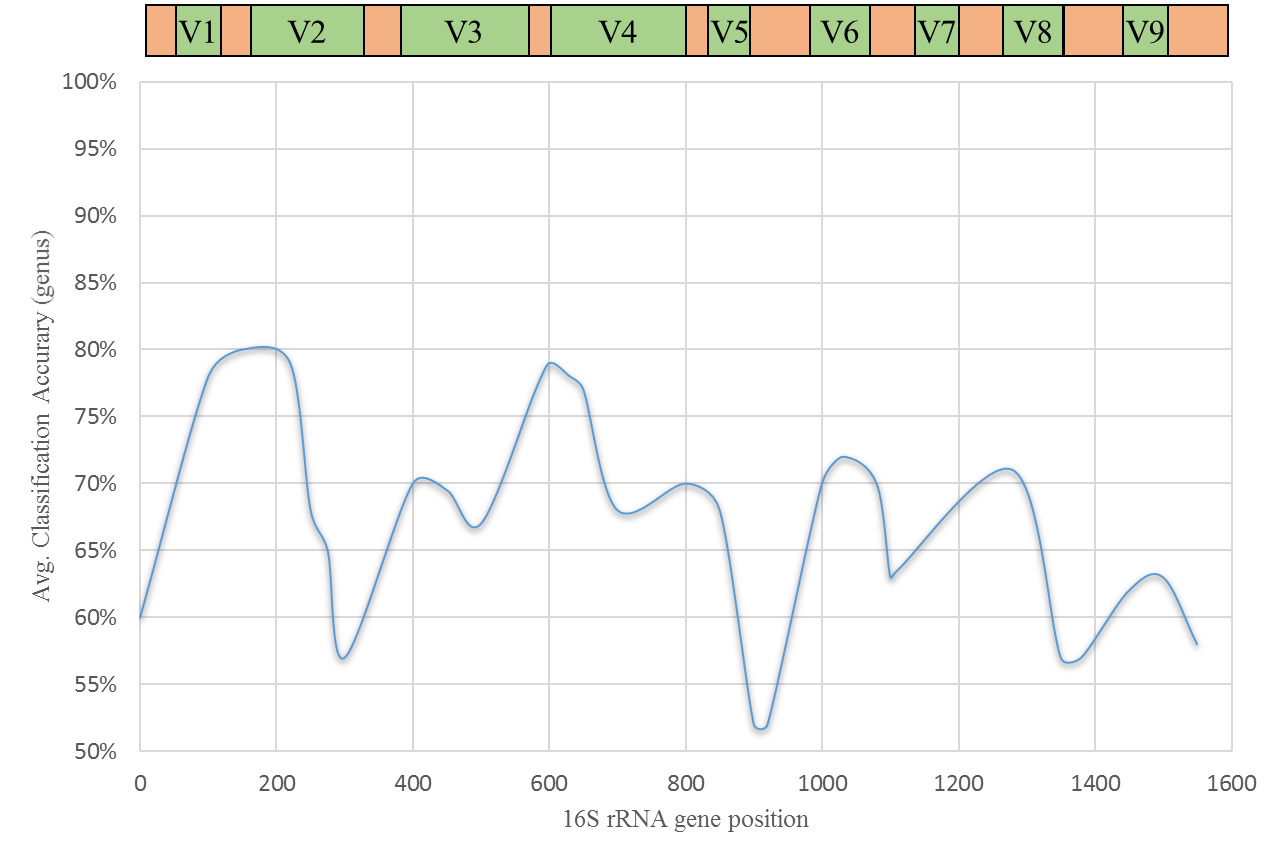


Figure .: Approximate representation of classification accuracy using a sliding average with variable regions denoted above the gene position. See Wang et al. for original work [[24](#_ENREF_24)].

As more sequences are identified through a numerous set of sequencing studies, previously unrecognized sequences will be classified against known sequences [[28](#_ENREF_28)]. Tools, such as QIIME and SILVA, help bin OTU’s taxonomically and more are in development [[28-30](#_ENREF_28)]. Some unique things that can arise out of these databases is a distinct OTU identifier that can allow for corroboration across multiple studies while others bin OTU’s into taxa but do not connect an OTU ID with a database [[29](#_ENREF_29)]. As tools omit an ID, relating findings to other literature prove challenging and most relationships only come in the form of phylum and on occasion a more in-depth taxonomic classification. Although ID’s are stand-ins for the original sequence, utilizing identifiers can only improve our knowledge of these organisms across a diverse set of environments.

## Metagenomics

Metagenomics is the study of all the genomes present in a microbial community as if it were one collective genome [[31](#_ENREF_31)]. Metagenomics has been an excellent tool in the realm of microbial communities because it has allowed the identification of potential machinery being used within a given environment and formulate hypotheses around mechanisms for survival or function of the community [[4](#_ENREF_4)]. Classical “shotgun” metagenomics focuses on sequencing long pieces of overlapping sequences such that these contiguous strands (contigs) can be parsed together and form an entire contiguous genome. Rather than focusing on unique sequences as with the 16S rRNA gene sequencing, metagenomes focus on identifying overlapping contigs at the sacrifice of more depth. While depth does play a role in identifying placement of contigs within the genome, identifying dynamics as they relate to microbial communities suffers. Therefore, the information more readily focuses on the machinery. With focus on the genes, taxonomic classification of where the genes are derived from become uncertain; methods such as MaxBin are designed to identify where the genes phylogenetically originate [[32](#_ENREF_32)].

With increased number of sequencing projects, many find themselves deposited in public databases providing excellent opportunities to create “meta” studies and provide new information from new bioinformatics tools. One such tool, PICRUSt, has evaluated a large database for known genes and related it back to taxonomic studies and OTU identifiers [[33](#_ENREF_33)]. PICRUSt, has unlocked new opportunities for 16S rRNA gene sequencing to gain additional insight into probable metagenomic profiles. This can be combined with knowledge about the environment to provide more comprehensive studies [[16](#_ENREF_16), [34](#_ENREF_34), [35](#_ENREF_35)].

Recent advances in metagenomics software has allowed for prediction of genomes from 16S rRNA gene sequences with the first of the software introduced being PICRUSt [[33](#_ENREF_33)]. PICRUSt is able to use OTU identifiers coupled with a database of predicted genes to matchup both individual predicted genomes and entire metagenomes. More software being introduced into the arena of predicted metagenomes areTax4Fun and PAPRICA [[36](#_ENREF_36), [37](#_ENREF_37)]. These methods leave no question as to where the predicted genes come from as it utilizes the 16S rRNA gene data as a source of information, but uncertainty does exist around the true presence of the genes in the metagenome. In spite of this uncertainty tools such as PICRUSt are gaining notoriety because of the valuable knowledge they are adding to 16S rRNA studies now having increased insight into functionality [[16](#_ENREF_16), [38](#_ENREF_38)]

## Network analysis for studying microbial ecology

Ecological interactions arise out of six distinct motifs (direction independent) pictured in Figure 1.1. These are thought to make up the whole of all microbial interactions which should be easily examined in a laboratory but considering its application to microbial communities, the scale of possible interactions quickly outpaces the ability to test [[6](#_ENREF_6)]. A workaround is to use the interactions recorded in microbial sequencing data (or any set of count data as this technique may pertain to metagenomic, proteomic, and other -omic analysis aiming to understand interactions) as a stand in for observation.

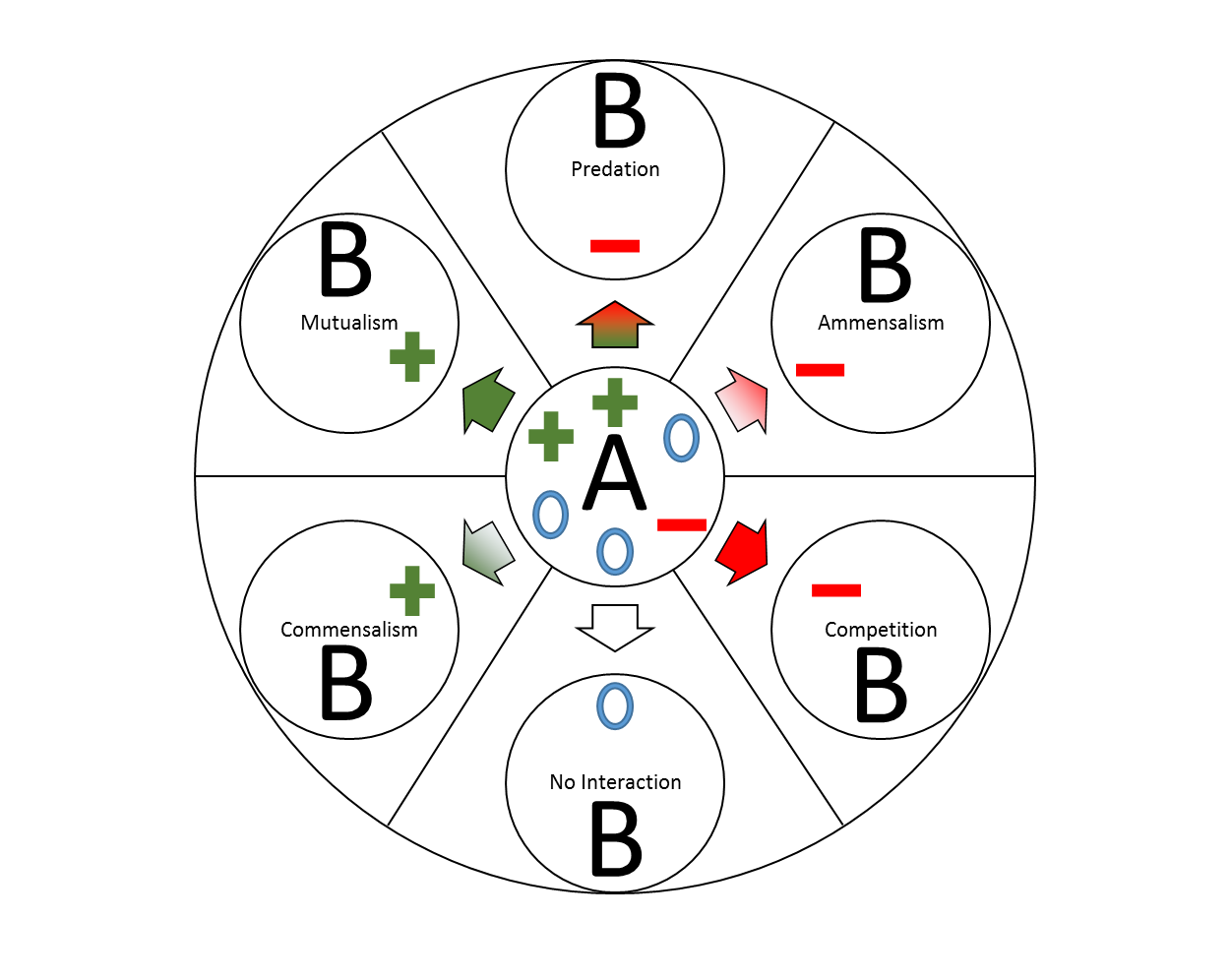


Figure .: All types of possible direction independent ecological interactions between hypothetical OTU’s A and B. The colors reinforce the sign for positive (green), negeative (red), and neutral (white/blue).

These networks can be constructed in a variety of ways but often aim to not only describe the strength and signedness of the interaction (Figure 1.2A), but also identify microbial niches (sub-communities) (Figure 1.2B) and relate the sub-communities to their working environment [[39](#_ENREF_39)]. Quantifying these sub-communities can come with its’ own complexity, but much can be borrowed from social network analysis. While community detection has not been optimized for sub-community detection of microbial niches, several methods have been applied [[40](#_ENREF_40)]. However, most methods discussed have not been rigorously examined across a wide array of possibilities as other social papers have examined [[19](#_ENREF_19), [41](#_ENREF_41), [42](#_ENREF_42)]. Methods such as Louvain, InfoMap, BoCluSt exist and this is an active area of research that tries to tease out unique niches that sub-communities fill [[43-45](#_ENREF_43)]. Identifying organisms that fill these sub-niches may be a critical element to improve prediction of microbial communities which is currently done by taxonomy [[46](#_ENREF_46)]. These other extensive comparison papers suggest other methods such a ‘InfoMap’ and ‘MarkovCluster’ although the ‘Louvain’ and ‘WalkTrap’ method followed closely. Each algorithm may in fact have its’ own merits based on network structure and size that may allow adaptable community detection methods [[20](#_ENREF_20)]. The identification of a community can further allow assessment of relationships between these sub-communities and environmental variables or outcomes [[47](#_ENREF_47)].

|  |  |
| --- | --- |
| A | B |

Figure .: (A) Networks are constructed based on connections between different OTU’s based on different correlation methods and the strength of the correlation is depicted by the thickness of the connection between the nodes (OTU’s). (B) Community detection methods will attempt to divide the network into smaller sub-communities as depicted by the different colors.

Studying these relationships can provide understanding into microbes that work together and the environment(s) they work in. Digging further into the operations of the microbes and their influence on the environment would be to understand potential cause-and-effect in the environment they are in. For example in the process of anaerobic digestion, it would be valuable to understand what sub-communities are responsible for acidogenesis, acetogenesis, and methanogenesis; using a process known as singular value decomposition (SVD) correlations can be drawn in between sub-communities and environmental variables [[47](#_ENREF_47)]. SVD is a common technique in matrix analysis and when applied to microbiome data treats OTU or gene count data as a matrix where sub-communities can be grouped together for further analysis. The SVD’s are correlated with the ebbs and flows of the respective environmental variable of interest to provide potential cause/effect as they relate to the sub-communities. This has provided new understanding of sub-communities and their effect on the environment [[26](#_ENREF_26)].

# Characterization of bacterial communities for high-solids deconstruction of lignocellulosic tomato processing waste

## Abstract

Lignocellulosic biomass is a proposed substrate for renewable energy that has encountered many technical hurdles to during the deconstruction process. Investigating food processing waste as a potential lignocellulosic biofuel substrate presents unique opportunities that would not only prevent carbon emissions associated with landfilling such waste, but to also displace the use of fossil fuels. Using a deconstructive microbial community can lead to more efficient understanding of the enzymatic needs of the industrial process. The characterization and predicted metagenome are presented here for compost-derived microbial communities deconstructing tomato pomace.

Using compost and water/EtOH extracted tomato pomace, bioreactors were set up to monitor CO2 and enzymatic activity of endoglucanases. DNA from multiple time points and environmental conditions were stored for 16S rRNA gene sequencing. 16S rRNA gene data was also combined with a software tool (PICRUSt) to reveal deeper insight into the potential function of the microbial metagenome. Measured endoglucanse activity saw a significant treatment effect of aeration and time. Community structure was influenced by treatment effects of aeration, time, temperature, and all interactions of those treatments. Aerated treatments saw increases in the presence of the phyla Proteobacteria and predicted ligninases. Non-aerated treatments saw dramatic increases in the phyla Firmicutes and predicted hemicellulases. The top five enriched OTU’s of each treatment accounted for a minimum of 40% of the mesophilic communities and 60% of the thermophilic communities. Ligninases and hemicellulases appear to be inversely correlated. Enriched enzymes were endo-1,4-beta-D-glucanase (K01179), alphafucosidase (K01206), and alpha-N-arabinofuranosidase (K01209). Pectinases were found to not make up a major portion of the metagenome of any treatment.

Environmental factors not only affect the community shape, but the function of the community as well. We show how not only specific environments interact but that the prevalence of certain types of genes can be inversely correlated to each other. Over time it was seen that metagenome profiles were less different suggesting some convergence in functionality. Overall, unique organisms and enzymes for the deconstruction of tomato pomace were identified.

## Introduction

The use of lignocellulosic biomass as a substrate for biofuel production has been a long-standing goal to decrease dependence on fossil fuel sources and mitigate carbon emissions [[48](#_ENREF_48)]. The use of food-processing wastes offers the opportunity to both mediate the disposal into landfills and provide a source of lignocellulose. Tomato processing waste, tomato pomace, offer approximately 20,000 Mg yr-1 of food-processing waste in California [[49](#_ENREF_49)]. However, deconstruction of this tomato pomace into fermentable sugars faces substantial technical hurdles as cellulose, the primary source of such sugars, is often shielded from enzymatic degradation within the cell wall via a matrix of lignin, hemicelluloses, and pectin [[50](#_ENREF_50)]. These components must be separated or degraded to make the cellulose accessible to hydrolytic enzymes such as endo-, exo-glucanases [[51](#_ENREF_51)]. As a result, the use of enzyme “cocktails” consisting of an array of deconstructive enzymes capable of degrading the various cell wall polymers have been suggested to increase the release of fermentable sugars from the cell wall matrix.

Cell wall degrading enzymes (CWDE) may be classified based on their target substrate within lignocellulose, such as cellulases, hemicellulases, ligninases, and pectinases. However, their activity on lignocellulose may also be influenced by the synergistic action of various other CWDEs. For example, enzymatic degradation of one cell wall component may improve accessibility for other enzymes to access their cell wall substrates. Synergies may also arise from enzymes that target different sites within a common substrate, such as different types of hemicellulases that preferentially cleave branches or main chain sites within a given hemicellulose polymer [[52](#_ENREF_52)]. Additionally, process variables such as feedstock type, temperature, and solids loading, can affect the relative activities of CWDEs [[53](#_ENREF_53)]. Given the complex interactions of CWDE mixtures, strategies are needed to inform design CWDE cocktails and maximize deconstruction activity on target feedstocks under industrially relevant conditions. To improve the efficacy of biofuels using CWDE mixtures, examining high-solids conditions are most germane [[54](#_ENREF_54)]. Lignocellulolytic microbial communities enriched from high-solids environments where lignocellulose degradation naturally occurs, such as in soil or compost, are a logical starting point for identifying deconstructive enzymes with potential synergies [[55](#_ENREF_55)].

The use of 16S rRNA gene sequencing via next-generation sequencing platforms, such as Illumina, has provided insight into changes in phylogenetic composition within lignocellulolytic microbial communities during enrichment culture [[10](#_ENREF_10), [55](#_ENREF_55)]. Benefits of 16S rRNA gene sequencing include the ability to achieve relatively deep coverage of sequencing while multiplexing many samples due to only a single gene being sequenced for each cell. In contrast, whole metagenome and metatranscriptome sequencing require more sequencing resources per sample since the entirety of each cell’s genomic or transcriptomic content is sequenced. While 16S rRNA gene sequencing allows rapid and highly parallel phylogenetic profiling of microbial communities, it does not provide information regarding gene content and, by extension, the functional potential of microbial communities on par with whole metagenome sequencing. As a result, tools have been developed to combine the benefits of both sequencing approaches. Phylogenetic Investigation of Communities by Reconstruction of Unobserved States (PICRUSt) is one such publicly available bioinformatics tool [[33](#_ENREF_33)]. PICRUSt cross-references operational taxonomic units (OTUs) detected via 16S rRNA gene sequencing of a microbial community with databases of sequenced genomes to reconstruct predicted metagenomes. Thus, 16S rRNA gene sequencing data can be used to infer the presence of target genes in microbial communities and predict microorganisms that may contribute to certain activities. This approach is particularly relevant to screening of lignocellulolytic microbial communities to identify deconstructive consortia and predict the deconstructive enzymes they employ [[34](#_ENREF_34)].

The goal of this research was to utilize 16S rRNA gene sequencing and bioinformatics tools to elucidate the phylogenetic and CWDE composition of compost microbial communities that were enriched to deconstruct tomato pomace under high-solids conditions. Enrichment cultures were conducted at multiple temperatures and aeration conditions to capture a variety of deconstructive communities derived from a common inoculum. These data can inform design of deconstructive microbial communities and enzyme mixtures for industrial bioprocessing applications.

## Materials and Methods

### Bioreactor Preparation and Operation

Mature green waste compost was obtained 2010 from a commercial composting facility in Zamora, CA that utilized municipal green waste (primarily lawn and yard clippings). Compost was air-dried outdoors and stored in plastic bins under ambient conditions (~25 °C) until use. Tomato pomace was collected from a commercial tomato paste processing facility in northern California in 2013. Pomace was solar-dried and then stored in plastic bins until use. To remove soluble compounds and enrich for lignocellulose as the primary carbon source in the tomato pomace, the pomace underwent extraction in a Soxhlet extractor using ethanol for 24 hours and then deionized water for 48 hours. The deionized water in the Soxhlet extractor was replaced after the first 24 hours of the water extraction. Tomato pomace was then dried in a vacuum oven to 5.4% dry-basis and stored in sealed bottles under ambient conditions until use.

Glass media bottles with a volume of 250 mL served as bioreactors. Bioreactors were loaded with 4 to 5 g (dry weight) of extracted pomace inoculated with mature compost (90% pomace and 10% compost, dry weight basis). Biomass was wetted to achieve high-solids conditions by adding M9 minimal medium (no carbon source added) to reach 300% dry basis moisture content (75% wet basis). During experiments, bioreactors were weighed daily and water was added to replace lost moisture and maintain the target moisture content. Bioreactors were incubated at either mesophilic (35 °C) or thermophilic (55 °C) temperatures for the duration of the enrichment culture process. Enrichment cultures lasted three weeks, which spanned the initial inoculation time point (T0), a transfer time point at week one (T1) where fresh batches of extracted pomace were inoculated with cultured biomass from bioreactors (10% inoculation by dry weight using methods described above), and a final time point two weeks after the transfer (T2). Samples were taken from the biomass for DNA and enzyme extractions at time points T1 and T2.

All bioreactors were connected to a MicroOxymax respirometry system (Columbus Instruments, Columbus, OH) to control bioreactor aeration and monitoring production of carbon dioxide, methane, and hydrogen gases. Sensors within the MicroOxymax system utilized infrared absorbance to quantify carbon dioxide and methane levels and an electrochemical cell to measure hydrogen gas levels in bioreactor effluent gas. Bioreactors were configured according to the manufacturer’s specifications for either aerobic or anaerobic operation. In brief, aerated bioreactors contained both inlet and outlet gas ports in the bioreactor lid through which the system periodically refreshed the bioreactor headspace with air. Anaerobic bioreactors contained only a single outlet port to allow escape of produced gases. A check valve (crack pressure 0.5 psi., Qosina Corp., Ronkonkoma, NY) was in-line with the outlet to prevent oxygen contamination from the surrounding air. All reactors had an empty 250 mL glass bottle installed in-line with the outlet line to act as a gas collector. The system sampled from gas collection vessels and measured target gas levels every 0.8-3.3 hours. For aerated reactors, the reactor headspace was flushed and replaced with ambient air in parallel with gas sampling.

### Enzyme Extraction and Activity Assay

Enzymes were extracted using a procedure described previously [[3](#_ENREF_3)] . The extraction buffer contained 50% w/w ethylene glycol, 0.1% tween 80, and 1% NaCl. 3g (fresh weight) of cultured biomass was added to 27g of extraction buffer and placed on a horizontal shaker at 150 rpm for one hour. Each sample was then vacuum-filtered through 0.22μm filter paper and then approximately 20mL of permeate was transferred to a centrifugal diafiltration unit with a 3,000 kDa molecular weight cut-off (Vivaproducts, Littleton, MA). Diafiltration units were centrifuged according to manufacturer instructions. Extracted enzymes were dialyzed by periodically adding 50mM sodium acetate buffer, pH 5 to the retentate chamber to replace the volume lost as permeate during centrifugation. Dialysis continued until the calculated ethylene glycol concentration in the retentate was below 5% w/w. After dialysis, samples were concentrated to 100X and stored at 4 °C.

Extracted and dialyzed enzymes were then assayed for endoglucanase activity by exposing them to carboxymethyl cellulose and monitoring glucose release via a colorimetric reducing sugar assay. The substrate used was a 2% w/w solution of carboxymethyl cellulose in 50mM sodium acetate, pH 5 buffer similar to [[15](#_ENREF_15)]. Enzymes and substrate were mixed at a 1:1 ratio by volume in a 96-well microplate for a total reaction volume of 80 μL and then incubated for 2 hours at either 35 or 55 °C to match the temperature applied during the corresponding enrichment culture for the sample. 80 μL of 3,5-dinitrosalicylic acid (DNS) reagent containing 1.4% DNS, 1.4% sodium hydroxide, 28% Rochelle salts, and 0.07% sodium sulfite, was then added to each reaction. Samples with DNS buffer were heated at 95 °C for 5 minutes to react the DNS reagent with reducing sugars and develop color. The absorbance at 540 nm was measured for each reaction using a spectrophotometer (BioTek Instruments Inc., Winooski VT). Glucose standard solutions at 0, 1, 2, 3, and 5 g L-1 were processed in parallel with samples. Resultant standard curves of A540nm versus glucose concentration were used to calculate reducing sugar content in enzyme-digested samples. The sugar released during the endoglucanase assay was normalized by the amount of wet material used during the enzyme extraction procedure and the moisture content to evaluate sugar released on a per gram dry weight basis.

### DNA Extraction and Sequencing of the16S rRNA gene V4 Hypervariable Region

DNA was purified from samples of compost inoculum and cultured pomace from time points T1 and T2 using the PowerSoil DNA Isolation Kit (Mo Bio Laboratories, Inc., Carlsbad, CA) according to the manufacturer’s instructions. Purified DNA was stored at -20 °C until use.

The hypervariable V4 region of the 16S rRNA gene was then amplified using 515F and 806R primers linked to the Illumina adapter of which contain the adapter compliment, barcode, pad and linker sequences as described previously in [[10](#_ENREF_10), [56](#_ENREF_56)].

### Processing and Analysis of Sequencing Reads

Raw sequencing reads were processed using the Quantitative Insights into Microbial Ecology (QIIME) toolkit [[29](#_ENREF_29)]. Reads were quality filtered using default values (r = 3; p = 0.75 total read length; q = 3; n =0; c = 0.005%) [[57](#_ENREF_57)]. Filtered reads were clustered into operational taxonomic units (OTU) with at least 97% sequence similarity [[29](#_ENREF_29)]. OTUs were phylogenetically binned by mapping the representative sequence of each OTU to the Greengenes database (Version 13.8) based on a similarity threshold of 97%. All OTU IDs reported are Greengenes OTU IDs. OTUs were then processed using the PICRUSt tool [[33](#_ENREF_33)]. The OTU counts were normalized by using marker genes via the *normalize\_by\_copy\_number.py*. The normalized 16S rDNA numbers were thenmapped to KEGG Orthologs (KO) using the PICRUSt *predict\_metagenomes.py* command. Using a list of previously identified deconstructive KO’s and including several additional target KO’s for lignocellulose deconstruction, the metagenome was filtered to just the KO’s of interest. Using the filtered list of KO’s, data was fed into the tool Statistical Analysis of Metagenomic Profiles (STAMP) for determining statistical significance of the representation of predicted KO’s in the metagenome [[58](#_ENREF_58)]. Benjamini-Hochberg and a p-value of ≤0.05 was used to test for statistical significance and correct for multiple tests. Following the analysis with STAMP, significant KO’s were examined using the *metagenome\_contributions.py* command to identify contributing OTU’s to significantly overrepresented KO’s of interest. Metagenome contributions were presented using the phyla that contribute at least 5% of the target KO.

### Data Analysis

CO2 gas data was acquired from the MicroOxymax and was analyzed using analysis of variance (ANOVA) to determine significance of final timepoints between treatments.

Endoglucanase enzyme assays were measured against a standard curve to determine sugar released during the hydrolysis process. The sugar released was normalized by both the moisture content of the sample and the amount loaded into the enzyme extraction process. ANOVA was performed on the resulting data.

Using the output of *normalize\_by\_copy\_number.py*, constrained correspondence analysis was completed in the VEGAN package in R (version 2.3.1). A model was built using categorical variables for aeration, time (T1, T2), and temperature (mesophilic, thermophilic). Significance was determined using *anova()* in VEGAN which mimics an ANOVA through a permutation test.

## Results

### Microbial respiration and endoglucanase activity

CO2 evolution was measured for bioreactors during the enrichment culture to monitor microbial activity. Table 2.1 shows the cumulative evolved CO2 for each enrichment culture treatment. The mesophilic aerated reactors produced less CO2 (p≤0.065) than the thermophilic reactors. No significant difference was detected for CO2 evolution between the two temperature treatments for the non-aerated reactors. After 1-week, the reactors had evolved 56% and 42% of the total CO2 in the mesophilic and thermophilic non-aerated reactors respectively. The aerated reactors achieved 75% and 79% of total evolved CO2 after one week.

|  |  |  |  |
| --- | --- | --- | --- |
|  | Time (hrs.) | CO2  Non-aerated  (μL g DW-1) | CO2  Aerated  (μL g DW-1) |
| Mesophilic | 483.91 | 1383 (1576) A | 161,484 (11,823) B |
| Thermophilic | 484.97 | 2627 (1277) A | 222,511 (42,877) B |

Table .: Cumulative CO2 respiration from mesophilic aerated, mesophilic non-aerated, thermophilic aerated, and thermophilic non-aerated. Different letters indicate significant differences between treatments. DW= dry weight

Endoglucanase activity was measured in enzyme extracts from each enrichment culture as an indicator of lignocellulose deconstructive activity. All samples noticed a decrease in endoglucanase activity peaked going from T1 to T2 (Figure 2.1). Aeration of samples at thermophilic and mesophilic conditions saw an effect of increased endoglucanase activity. Measured endoglucanase activity noticed decrease in variation for mesophilic anaerobic conditions.

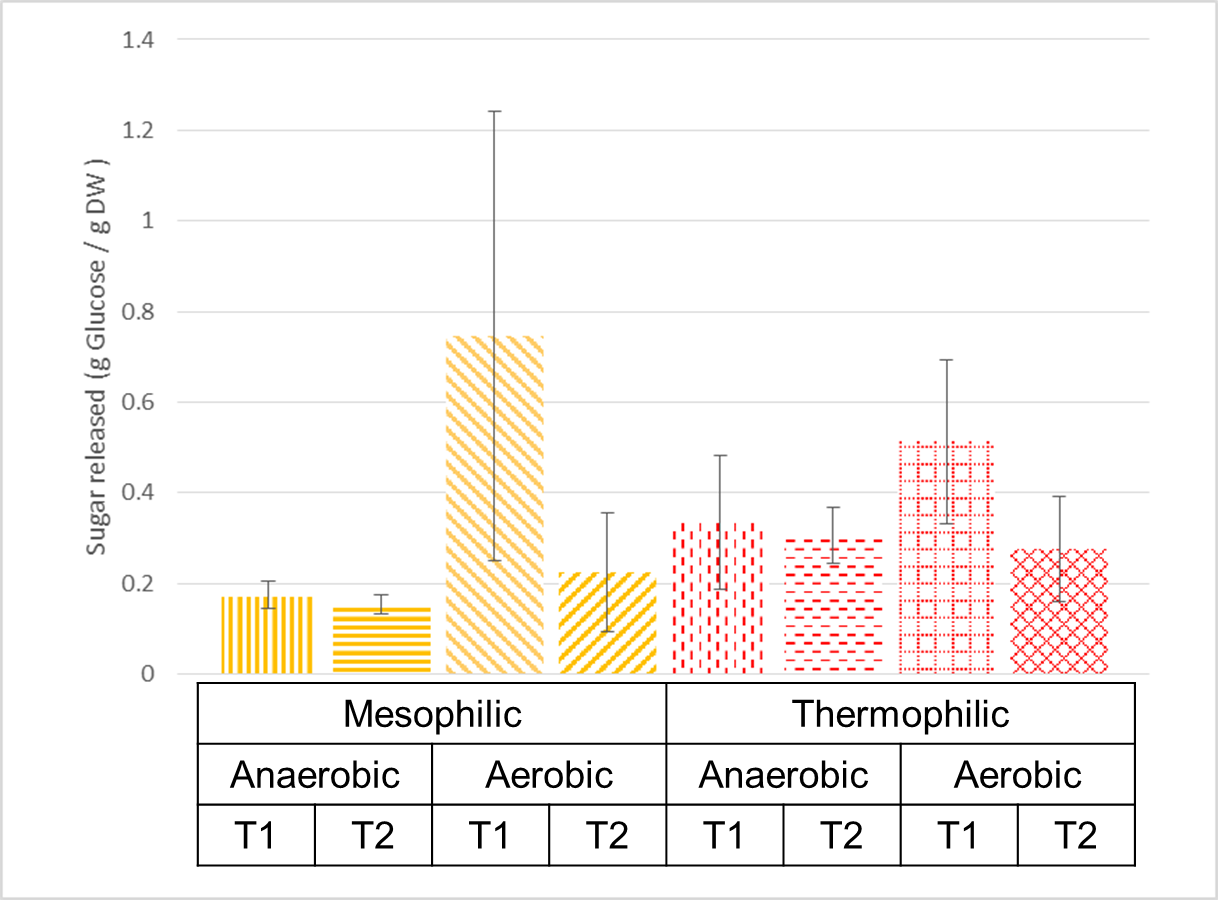


Figure .: Sugars released as measured by DNS assay from extracted enzymes on CMC.

Conducting an ANOVA on the released sugar, aeration and time had an effect on the enyzmes ability to degrade carboxymethyl cellulose (CMC) (Table 2.2). At T2, extracted enzymes released less sugar. This decrease in activity affected the aerated enzymes more significantly as they experienced a sharper drop in activity than that of the non-aerated extracted enzymes. Temperature had less significant (p>0.05) of an effect on the measured activity of the enzymes and was excluded from the model.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | Df | Sum Sq | Mean Sq | F value | Pr(>F) |
| Timepoint | 1 | 0.2461 | 0.2461 | 5.7105 | 0.0274 |
| Aeration | 1 | 0.2601 | 0.2601 | 6.0350 | 0.0238 |
| Timepoint x Aeration | 1 | 0.1649 | 0.1649 | 3.8253 | 0.0653 |
| Residuals | 19 | 0.8189 | 0.0431 |  |  |

Table .: ANOVA table of sugars released during the hydrolysis of Carboxymethyl cellulose (CMC) from extracted enzymes

### Taxonomy

Canonical correspondence analysis (CCA) was performed and showed differences between the communities (Figure 2.2). All communities shifted away from the initial starting microbial community. T1 communities were always more dissimilar than T2 as compared to the initial compost inoculum. Thermophilic communities followed a distinct trajectory perpendicular to that of the mesophilic communities. Non-aerated communities for both temperature regimes were more dissimilar than the aerated counterpart. The CCA was complemented with a permutation test to examine residual differences in building a constrained model (Table 3). The resultant analysis concluded that a full model (Table 3) with time, temperature, aeration regime and including the all interactions, were significant in describing shifts in the communities.

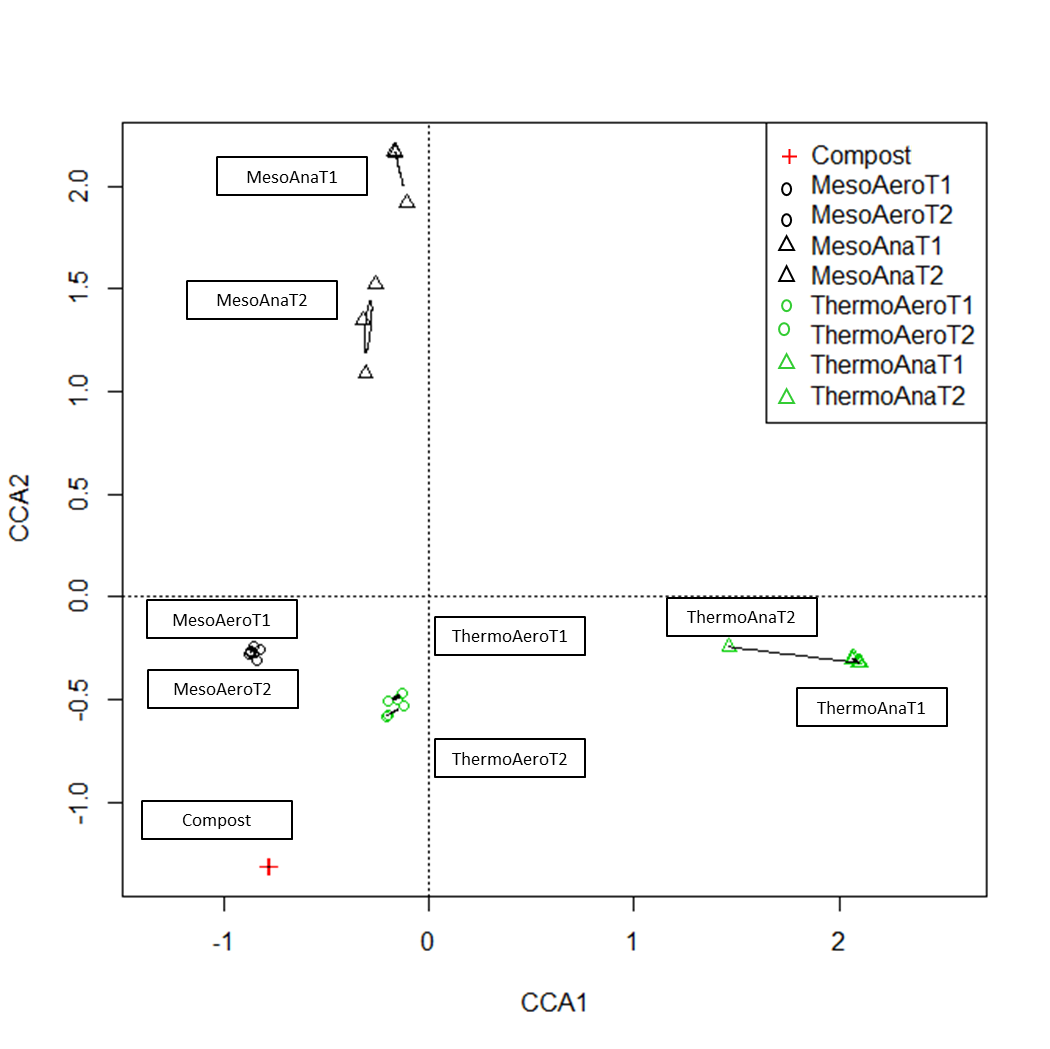


Figure .: Plot of Constrained Correspondence Analysis.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | Df | ChiSquare | F value | Pr(>F) |
| Aeration | 1 | 0.44769 | 7.6249 | 0.001 |
| Time | 1 | 0.18721 | 3.1884 | 0.001 |
| Temperature | 1 | 0.45014 | 7.6666 | 0.001 |
| Aeration x Time | 1 | 0.19524 | 3.3252 | 0.001 |
| Aeration x Temperature | 1 | 0.43020 | 7.3270 | 0.001 |
| Time x Temperature | 1 | 0.17089 | 2.9104 | 0.001 |
| Aeration x Time x Temperature | 1 | 0.16703 | 2.8448 | 0.002 |
| Residual | 15 | 0.88073 |  |  |

Table .: ANOVA table from the result of a permutation test on the Constrained Correspondence Analysis from the OTU’s present in each samples

Despite being most similar to the inoculum, the aerated communities show dramatic shifts in relative abundance of the phyla present (Figure 2.3).). Most noticeable is the increase in the relative abundance of the phyla Proteobacteria with 81.5% and 76.5% being present in the T2 mesophilic and thermophilic community respectively. This large increase also was corresponded with a decrease in the initial relative abundance of Actinobacteria from 29.7% to 2.9% and 5.2% for T2 mesophilic and thermophilic. More distinctly the phyla Chloroflexi decreased from 40.0% to 0.2% and 2.2%. There is also a noticeable increase in the phyla Gemmatimonadetes from the inoculum (1.4%) that is more present in the T1 (3.5% and 14.4%) than T2 (2.4% and 4.3%). Verrucomicrobia saw a small relative abundance at T2 of the mesophilic aerated condition (2.4%) but did not seem to represent any large portion of any other condition (<0.5%).

The non-aerated communities saw the greatest shifts of Firmicutes from the inoculum (3.8%) with large increases in the relative at both T1 (93.8% and 99.4%) and T2 (56.7% and 97.8%) of the phyla within the mesophilic and thermophilic conditions. Non-aerated communities also noticed large decreases of Actinobacteria and Chloroflexi from the initial compost. AA Small increase of Proteobacteria was noted for the mesophilic T2 non-aerated (30.6%) from the initial 15.1% present in the inoculum. The mesophilic non-aerated community saw the only increase in the phyla Tenericutes for T1 (5.0%) and T2 (9.8%).

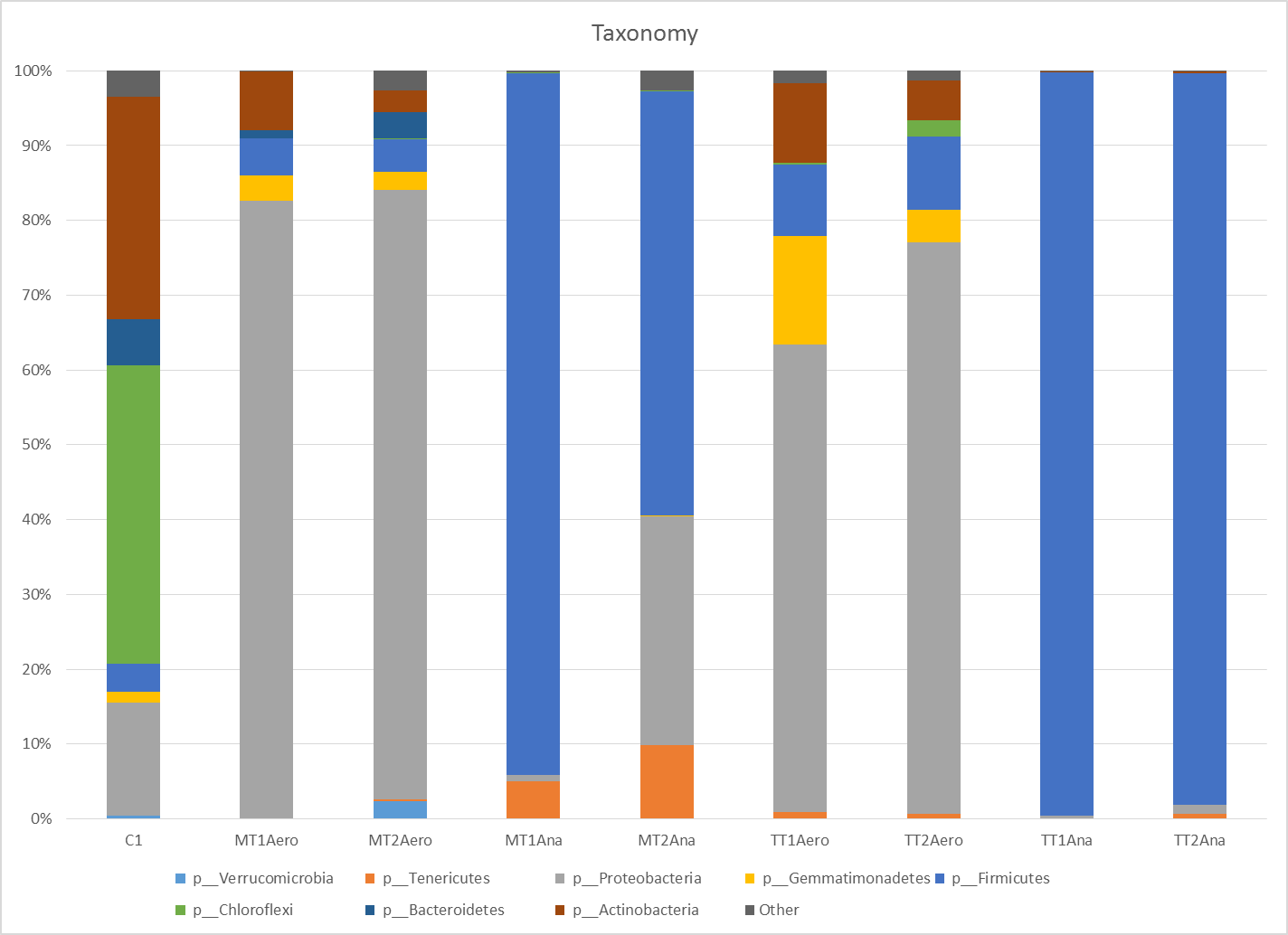


Figure .: Relative abundances of the initial, T1, and T2 communities. “p\_” indicates we are referring to the Phyla. Phyla present in all communities at greater than 2% are shown whereas the remaining phyla are classified as “Other”.

The dramatic shifts in OTU’s present within the communities could be attributed to being dominated by a select number of organisms (Table 2.4). The top five OTU’s for mesophilic communities could describe 40% - 50% of the total community whereas those numbers increased to 70% - 80% for thermophilic conditions. While the non-aerated conditions were dominated by the phylum *Firmicutes*, the mesophilic condition showed a bias away from this phylum for the most abundant OTU’s. Aerated conditions derived the most abundant OTU’s from the phylum *Proteobacteria*, but had a minor contribution from *Firmicutes* within the thermophilic conditions.

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| MT2Aero |  | MT2Ana |  | TT2Aero |  | TT2Ana |  |
| 691349 (P, *Sinobacteraceae*) | 23.1% | 1097359 (P, *Acinetobacter lwoffii*) | 22.5% | 513055  (P, *Steroidobacter*) | 35.5% | 565764 (F, *Ruminococcus*) | 46.1% |
| 587540 (P, *Sinobacteraceae*) | 9.4% | 4305692 (T, *ML615J-28*) | 5.0% | 648242 (P, *Halorhodospira*) | 12.5% | 749791 (F, *Ethanoligenens*) | 11.4% |
| 636316 (P, *Sphingomonadaceae*) | 8.7% | 4459356 (F, *Ureibacillus*) | 4.7% | 365415 (P, *Steroidobacter*) | 11.8% | 257863 (F, *Thermoanaerobacterium*) | 6.4% |
| 998905 (P, *Mycoplana*) | 4.3% | 557051 (F, *Ruminococcaceae*) | 4.4% | 2700470 (P, *Steroidobacter*) | 7.9% | 578347 (F, *Planococcaceae*) | 4.4% |
| 1108275 (P, *Limnohabitans*) | 3.8% | 246274 (F, *OPB54*) | 4.3% | 549672 (F, *OPB54*) | 2.4%% | 130040 (F, *Alicyclobacillus*) | 2.9% |

Table .: Top five OTU relative abundances from the final time point of each community. Numbers indicate the Greengenes ID, single letter represents the phylum (F – Firmicutes, P – Proteobacteria, T – Tenericutes), and italicized taxonomy represents the highest degree of discernible classification obtainable using 97% similarity.

### Metagenome

Shifts in the phylogenetic composition of the communities is directly related to shifts in the metagenome and deconstructive capacity of the community. 45 genes were analyzed within the community that were related to plant cell wall deconstruction. A heatmap shows the relative abundance of genes within each reactor (Figure 4). A greater abundance of hemicellulase enzymes (cross-linking glycans) were present in the metagenome of the non-aerated communities than the aerated conditions. The dominant hemicellulases in the non-aerated conditions were K01206 (alpha-L-fucosidase) and K01209 (alpha-N-arabinofuranosidase) Likewise, lignin degrading enzymes represented a larger portion of the aerated metagenome. The mesophilic aerated samples had high levels of the K03862 (vanillate monooxygenase) and K03386 (alkyl hydroperoxide reductase subunit C). The thermophilic lignin enzymes were again K03386 and K00432 (glutathione peroxidase). Pectin degrading enzymes (K01051 and K01194) were present but only in low quantities. Cellulose degrading enzymes were predominantly K01179 (endoglucanase) and K05349 (beta-glucosidase).

Using a combination of statistical analysis of metagenomic profiles (STAMP) and principal component analysis (PCA), genes were analyzed for statistical significance by aeration or temperature and then projected into the first to major principal components. This was done for each time point. During T1 the first two PCA axes accounted for 78.4% of the variation while at T2 the axes accounted for 76.1%. During T1, 22 genes were found significant due to aeration effects and one to temperature. T2 had only six genes that were significant all because of aeration. Figure 2.4AA shows a significant distinction between the direction of much of the lignin degrading enzymes and hemicellulose degrading enzymes as they appear to be inversely correlated. The lignin degrading enzymes point towards the direction of the aerated treatments and hemicellulose degrading enzymes towards the non-aerated treatments. Figure 2.4AA also shows that two cellulose degrading genes (K01179, endo-1,4-beta-glucosidase; K05349, beta-glucosidase) were correlated with the hemicellulose degrading genes. Only one pectin enzyme was found to be significant but did not appear to be correlated with either lignin or hemicellulose degrading genes but was associated with mesophilic aerated conditions.

Figure 2.4B showed a drastic change in metagenome makeup with a decrease in the number of significant genes and two enzyme classes (cellulose, pectin) that were no longer significant. While lignin degrading enzymes still correlated with one another, only one hemicellulose degrading enzyme was inversely correlated while another projected orthogonal to the lignin degrading enzymes towards mesophilic aerated conditions. This enzyme (K01224), was previously associated between both thermophilic and mesophilic aerated conditions.

The enzymes appear to group less around a single type of treatment as the axes show less difference between the treatments themselves (excluding thermophilic non-aerated community). The thermophilic non-aerated community saw a drastic shift away from the other treatments.

|  |
| --- |
| A |
| B |

Figure .: Principal Component Analysis of significant genes at T1 (A) or T2 (B).

|  |  |
| --- | --- |
| A | B |
| C | D |

Figure .: Relative abundance of target genes on target substrates for compost, T1, and T2.

#### Cellulase

##### Aerated

Examining cellulose deconstructive enzymes within the metagenome (Figure 2.5A), dramatic increases in Proteobacteria were detected in the aerated condition contributing 59.6% of the mesophilic cellulose metagenome and 41.3% in the thermophilic conditions. Proteobacteria predominantly added to the presence of endoglucanases (K01179, K01188) up to 48% of the respected ortholog in an individual reactor. Relative to the green waste compost, the aerated conditions also derived more of their cellulose metagenome (K01179) from Gemmatimonadetes. Actinobacteria made up a large portion of all aerated conditions while decreasing from T1 to T2. Chloroflexi noticed dramatic decreases in relative abundance of contributing genes to the cellulase metagenome ending at 0.2% (mesophilic) and 7.1% (thermophilic). Verrucomicrobia also saw a small contribution (6.8%) to the mesophilic T2 aerated condition but no other major contribution. The T2 mesophilic treatment saw an increase in the phyla Bacteroidetes that brought an increase in endoglucanases and beta-glucosidases. The increase could be attributed to three distinct OTU IDs: 891031 (*Sphingobacterium multivorum*), 191237 (*Chryseobacterium*), and 3511168 (*Cytophagaceae*). *S. multivorum* and *Chryseobacterium* was predicted to only having beta-glucosidases within the category of cellulose deconstruction. *Cytophagaceae* also had an endoglucanase (K01179).

Cellulase genes that were significantly overrepresented during T1 were K05349 (beta-glucosidase), K01225 (1,4-beta cellobiosidase) but none were detected during T2. Within the top five OTU’s, aerated samples saw the presence of at least one cellulase gene with the exception of *Limnohabitans*. K01179 was the most common to encounter among the most abundant OTU’s, followed by K05349. Within the abundant OTU’s, K01187 and K01188 (OTU ID’s: 636316 and 998905) were only detected for the mesophilic whereas only K01225 and K05350 were only detected among thermophilic conditions (OTU ID’s 549672). *Sphingomonadaceae* was identified as having the most number of unique cellulases (K01179, K01187, K01188, K05349) under the mesophilic abundant OTU’s and *OPB54* (K01179, K01225, K05349, K05350) under the thermophilic abundant OTU’s.

##### Non-Aerated

Non-aerated conditions were almost completely dominated by Firmicutes within the cellulase metagenome (Figure 2.5). The mesophilic condition saw the largest contribution of K01179 genes and K05349 from the genus *Ruminococcus* (OTU ID: 1089636) within the Firmicutes phylum. At thermophilic conditions, *Ruminococcus* still made the largest contribution of the aforementioned enzymes, but was identified as OTU ID: 565764. Thermophilic conditions also present the genus *Ethanoligenens* (OTU ID: 749791) as a major contributor of endoglucanase (K01179, K01188) and beta-glucosidase (K05349, K05350). Small contributions were also made by the phyla Tenericutes (mesophilic T2, 3.6%) and Proteobacteria (mesophilic T2, 4.1% and thermophilic T2, 1.0%). Tenericutes was identified as only having K01188 (endoglucanase).

No detected cellulases were found for *Acinetobacter lwoffii* despite being the single most abundant OTU in the mesophilic community. Within Table 4 of the non-aerated community *Ureibacillus* had the greatest number of unique cellulases for the mesophilic community (K01179, K01187, K01188, and K05350) and *Ruminococcus ethanoligens* had the most unique set for thermophilic conditions (K01179, K01187, K01188, K05349, K05350).

#### Hemicellulase

##### Aerated

Proteobacteria represented a disproportionate amount of the hemicellulose metagenome as compared to the taxonomy as mesophilic T2 represented only 43.6% and thermophilic T2 only 10.8%. The most dominant OTU’s for mesophilic Proteobacteria was *Sphingomonadaceae* (OTU ID: 636316). This was distinctly different from the most dominant thermophilic *Myxococcales* (OTU ID: 733046). By mesophilic T2, the phyla Verrucomicrobia and Bacteroidetes had an increase in presence within the hemicellulose metagenome to 14.5% and 12.4% respectively. This increase was due to an increase in *Opitutus* (OTU ID: 4299150) and *S. multivorum* (OTU ID: 891031) for Verrucomicrobia and Bacteroidetes respectively. Most notable were the thermophilic aerated conditions that contained a large representation of Chloroflexi hemicellulase genes (18.3%) within the metagenome, which were of greater relative abundance than the Chloroflexi in the actual community (2.2%). While Chloroflexi was enriched between T1 and T2 thermophilic, Gemmatimonadetes saw a decrease in the relative abundance of hemicellulase genes (22.2% to 8.8%).

Among the most abundant OTU’s, a large proportion were not identified as having any of the 25 hemicellulase genes. For the mesophilic conditions, only *Sphingomonadaceae*, *Mycoplana*, and *Limnohabitans* contained any of the genes and *Limnohabitans* was only predicted to have K03927 (carboxylesterase 2). Under thermophilic conditions only *OPB54* was identified as having hemicellulase genes but contained eight unique genes (K01181, K01190, K01192, K01198, K01206, K01209, K07406, K12308).

##### Non-aerated

For the non-aerated conditions, Firmicutes represented the largest proportion of genes, 90.9% (mesophilic T2) and 99.3% (thermophilic T2) contributing to hemicellulases within the metagenome. The largest contribution to the hemicellulase came from *Ruminococcus* for both mesophilic and thermophilic (OTU ID: 1089636 and 565764 respectively). The dominant genes were K01206 (alpha-L-fucosidase) and K01209 (alpha-N-arabinofuranosidase). Under thermophilic conditions, *Ethanoligenens* (OTU ID: 749791) was observed just as with the cellulases. The small increases in the presence of Proteobacteria during mesophilic (6.6%) and thermophilic (1.5%) T2 was due to OTU ID’s: 1097359 (*Acinetobacter lwoffii*) and 733046 (*Myxococcales*) respectively. Despite contributing to the cellulase metagenome, Tenericutes was not detected during the PICRUSt reconstruction of genes encoding for hemicellulases.

Even though *Acinetobacter lwoffii* was noted for its increase in the hemicellulase metagenome, the only hemicellulase gene identified within its predicted metagenome was K01044 (carboxylesterase 1). Among the abundant OTU’s the mesophilic conditions had one OTU, *ML615J-28*, identified as not having any hemicellulase genes under whereas all of the abundant OTU’s for thermophilic conditions were identified as having at least on hemicellulase gene. Large contribution came from *Ruminococcus* (OTU ID: 565764) as a result of both abundance (46.1%) and the 11 unique hemicellulase genes. Under thermophilic conditions, this OTU and *Thermoanaerobacterium* were identified as having 11 unique hemicellulase genes with the next closest OTU (*Ethanoligens*) having nine. The largest set of unique genes for mesophilic conditions was *Mycoplana* and *Sphingomonadaceae* with five.

#### Pectin

##### Aerated

While the inoculum was dominated by Actinobacteria (58.6%), the mesophilic aerated community enriched for Bacteroidetes (47.0%) as a source of Pectin degrading enzymes for the community. The main OTU ID for Bacteroidetes was 891031 (*Sphingobacterium multivorum*) but only represented 2.0% of the enriched community. The second most dominant phylum was Verrucomicrobia which contributed 28.7% mainly due to the OTU ID: 4299150 (2.4% relative abundance in taxonomy). However, at thermophilic conditions, Bacteroidetes did not represent more than 0.4% of the relative abundance of genes. At thermophilic aerated conditions, Actinobacteria and Firmicutes appeared to be the dominant contributing members to the pectin degrading metagenome. The leading contributor from Actinobacteria was *Thermobispora bispora* (OTU ID: 689814) whereas for Firmicutes it was *Panibacillaceae* (OTU ID: 885978). The majority of the metagenome for pectin degradation came from K01051 (pectinesterase). It should be noted that the relative abundance of pectin genes within the entire metagenome represented a very small fraction as pictured in Figure 4.

##### Non-Aerated

Non-aerated conditions were mostly dominated by Firmicutes except at T2 thermophilic as, Actinobacteria were enriched to contribute 1.8% more genes encoding for pectin degrading enzymes. The dominant member under the mesophilic conditions was *Enterococcus casseliflavus* (OTU ID: 584241) and *Thermobispora bispora* (OTU ID: 689814) under thermophilic conditions. These OTU’s were otherwise only small contributions to the overall taxonomy. Again, K01051 was the dominant gene within the pectin degrading metagenome. Despite these phyla representing a major portion of the relative abundance, Tenericutes, Gemmatimonadetes, and Chloroflexi, were not detected amongst the pectin degrading enzymes within the reconstructed metagenome and none of the most abundant OTU’s were identified as having any pectinases.

#### Lignin

##### Aerated

For lignin degrading enzymes, this enriched for Proteobacteria across all treatments (Figure 6D). For aerated treatments, mesophilic conditions saw the presence of Bacteroidetes (4.9%) and Verrucomicrobia (0.8%), but at thermophilic temperatures were not present (<.01%). Two of the most common genes in the mesophilic temperatures were K00432 (glutathione peroxidase) and K03862 (vanillate monooxygenase). These were vastly contributed by OTU ID’s 636316 (*Sphingomonadaceae*) and 691349 (*Sinobacteraceae*). At thermophilic temperatures the most common genes were K00432 and K0433 (non-heme chloroperoxidase) and were largely contributed by OTU ID 513055 (*Steroidobacter*).

While *Limnohabitans* was predicted to only have one other deconstructive gene, it represented the greatest number (K00104, K00428, K00432, K03386, K03781, K03862, K03863) of unique lignin degrading enzymes. All of the abundant mesophilic and thermophilic OTU’s contained multiple ligninases except for *OPB54* which only contained one (K00104, glycolate oxidase). The minimum number of unique ligninases from the abundant OTU’s was four under mesophilic conditions and, excluding *OPB54*, four for thermophilic conditions.

##### Non-Aerated

During non-aerated conditions, Firmicutes represented a major portion of the lignin degrading enzymes in the reconstructed metagenome. However, an enrichment of Proteobacteria was seen during the mesophilic conditions and represented 51.6% of the lignin metagenome. Tenericutes was not detected in any of the samples but did see a minor presence in the original inoculum. Most of the mesophilic lignin metagenome from Proteobacteria was derived from OTU ID: 1097359 (*Acinetobacter lwoffii*). This contributed to K03781 (catalase) being the most abundant gene. Under thermophilic conditions, a OTU ID: 749791 (*Ethanoligens*) had the largest contribution to the lignin metagenome. The most prevalent gene under thermophilic conditions was K03386 (peroxiredoxin).

While ligninases were not as common as within the aerobic lignin metagenome, most of the abundant OTU’s contained some form of a lignin degrading enzyme. Specifically, *Acinetobacter lwoffii* contained six unique (K00432, K0433, K03386, K03781, K03862, K03863) ligninases. *ML615J-28*, *Ruminococcaceae*,and *Ruminoccocus* from the most abundant OTU’s were not predicted to have any ligninases. The most common type of gene to have amongst the abundant OTU non-aerated ligninases was K03386 for both mesophilic and thermophilic conditions.

## Discussion

Microbial enrichment of compost-derived communities saw dramatic shifts in community composition due to temperature, aeration, and time that all will affect both the community structure and metagenome. Despite aerated and non-aerated producing different levels of CO2, similar enzyme activities were detected for endoglucanases at T2. This may have to do with the production of non-CO2 metabolites such as acetate, propionate, and butyrate often produced through anaerobic pathways [[59](#_ENREF_59)]. These metabolites can make up a significant portion of metabolite production during the deconstruction process. Though some of these organic acids may be recoverable as energy by some organisms [[60](#_ENREF_60)], they can also accumulate and act as inhibitors [[61](#_ENREF_61)]. It is also possible there is decreased metabolic activity in the anaerobic conditions due to limitations surrounding substrate access. This would largely appear from lignin or lignin degradation by-products [[62](#_ENREF_62)]. Similar would be inhibition due to the toxicity associated with lignin during the deconstruction process [[11](#_ENREF_11)]. Microbial enrichment of compost-derived communities saw dramatic shifts in community composition due to temperature, aeration, and time that all will affect both the community structure and metagenome. Despite aerated and non-aerated producing different levels of CO2, similar enzyme activities were detected for endoglucanases at T2. This may have to do with the production of non-CO2 metabolites such as acetate, propionate, and butyrate often produced through anaerobic pathways [[59](#_ENREF_59)]. These metabolites can make up a significant portion of metabolite production during the deconstruction process. Though some of these organic acids may be recoverable as energy by some organisms [[60](#_ENREF_60)], they can also accumulate and act as inhibitors [[61](#_ENREF_61)]. It is also possible there is decreased metabolic activity in the anaerobic conditions due to limitations surrounding substrate access. This would largely appear from lignin or lignin degradation by-products [[62](#_ENREF_62)]. Similar would be inhibition due to the toxicity associated with lignin during the deconstruction process [[11](#_ENREF_11)].

Endoglucanase activity as measured through the DNS assay saw changes over time with no significant differences between all reactors at T2. Though activities may not have been significantly different at T2, Table 2.2 shows that both time and aeration impact the sugars released due to enzymatic degradation. However, over time the activity of enzymes from the aerated communities decreased and this may have to do with the depleted nitrogen, degradation of enzymes, or accumulated inhibitors among many possibilities. It was also noted that no significant differences in cellulase genes were detected in the metagenome at T2. The metagenome suggests that most of the genetic enrichment in the aerated communities targeted lignin deconstruction (Figure 2.4).). This may have to do with decreased access cellulose chains or lignin inhibition of enzymes as the tomato pomace undergoes deconstruction [[62](#_ENREF_62)].

As lignin degradation is largely considered an aerobic metabolic activity, this may explain the divergent enrichment between the non-aerated communities and the similar endoglucanase activity as the communities diverged to contain ligninases and hemicellulases. While ligninases were detected in the non-aerated community, it is unknown whether or not the community was utilizing these genes and whether they were in active form. Lucey and Leadbetter identified a lignin monomer degradation pathway in an anaerobic bacterium, but the enzyme was non-functional without oxygen [[63](#_ENREF_63)]. Lucey and Leadbetter however did suggest that the anaerobic lignin degrading enzymes may be a source of antioxidant genes that consume oxygen through phenolic ring opening to prevent oxygen contamination. Knowledge surrounding anaerobic degradation of lignin is particularly limited as oxygen-dependent enzymes are usually involved [[64](#_ENREF_64)].

As the majority of respiration occurred during the first week, the community shape was most likely influenced by the recalcitrant polymers during the later stages of the enrichment as substrate is known to influence community structure [[55](#_ENREF_55)]. This can be seen by the shifts in the community (Figure 2.3) and subsequent metagenome (Figure 2.5). Utilizing the relative abundance of deconstructive genes and their respective functions, it is inferred that pectin did not play a major role in shaping the community as the relative abundance (Figure 2.5) for pectin degradation and showed deviation from overall community relative abundance (Figure 2.3) and low overall presence within community metagenomes. It is hypothesized that even though tomato pomace is known for being pectin-rich that pectin was extracted during the EtOH and hot water Soxhlet extraction process. The PCA of the deconstructive genes (Figure 2.4) suggests how certain genes can be enriched in target conditions although the decreasing number of significant genes suggest that most genes approach a steady-state within overall plant cell wall deconstruction.

Most of the target enzymes of interest were identified among just several of the most dominant OTUs (supplementary file X) which is consistent with previous deconstruction work [[65](#_ENREF_65)]. Temperature affected metagenomes less than aeration suggesting that mostly convergent functionality with different communities within aeration methods. While differences in the presence of endoglucanase within the class of enzymes for cellulose deconstruction were found (supplementary file X), they did not appear to associate with measured endoglucanase activities (Figure 2.1). This could be influenced by factors such as different activity levels of the secreted endoglucanase, transcription levels, or pH of enzymatic environment although it has been shown previously that there is high correlation (R2 = 0.895) between gene abundance and enzyme activity of laccase multicopper oxidase during a composting process [[13](#_ENREF_13)]. However, there could be disconnect between the two types of genes as the PCA of the deconstruction metagenome does show inverse relationships between the endoglucanase (K01179) and most ligninases while appearing to align with most hemicellulases (Figure 5).

While most hemicellulases appear to linked to the non-aerated environment (Figure 5A) and also represent a significant portion of the genomes for the most abundant OTU’s (supplementary file X), *Acinetobacter lwoffii* was able to thrive as a lignin degrading bacteria in this environment which may have provided synergistic enzyme activities [[18](#_ENREF_18)]. Although it is hypothesized that these synergistic effects may not be the dominating type of microbial interaction [[6](#_ENREF_6)]. This seems to corroborate that the majority of the most abundant OTU’s tend to exhibit genetic redundancy and thereby most likely similar metabolism and competition for like resources (substrate). The enrichment of the most common hemicellulases (alpha-L-fucosidase and alpha-N-furanosidase) has been seen previously during the deconstruction of wheat straw and poplar wood chips [[34](#_ENREF_34), [65](#_ENREF_65)]. However, during the anaerobic deconstruction of the poplar wood chips, these enzymes were found within *Bacteroides* [[65](#_ENREF_65)]. The enrichment of *Firmicutes* may have to do with minor differentiation of hemicelluloses found within tomato pomace and preferences for hemicellulose degradation as wheat straw has a different sugar composition [[66-68](#_ENREF_66)].

The prevalence of ligninases in the aerated environment but not in the non-aerated environment is nothing new, so the deficit of ligninases was not surprising, but the metagenome highlighted distinct differences between the types of specific lignin-type activity. The uniqueness of *Acinetobacter lwoffii*, although not previously seen with this species, to degrade lignin anaerobically has been seen prior within the *Acinetobacter* genus and may be beneficial to the community [[69](#_ENREF_69)]. It has been suggested that other *Acinetobacter* species play this critical role in lignin degradation for deconstructive communities [[16](#_ENREF_16)]. Typically, the non-aerated communities were lacking five out of the eight common ligninases. This may drive the divergence towards the non-aerated communities accumulating bacteria with alternative cell-wall degrading enzymes such as hemicellulases despite hemicellulose only making up 8% - 15% of tomato pomace [[67](#_ENREF_67)].

Despite tomato pomace being lignocellulosic material, the composition of the various metagenomes diverged due to aeration and therefore did not match the composition of the tomato pomace suggesting that plant cell wall polymers are degraded non-arbitrarily [[67](#_ENREF_67)]. Aerated communities were enriched for specific aromatic degrading enzymes (glutathione peroxidase, vanillate monoxygenase, and non-heme chloroperoxidase), that may have been critical to the access of cellulose polymers as lignin represents 33% - 37% of the tomato pomace [[67](#_ENREF_67)]. This heavily influenced the aerated community as *Proteobacteria* were heavily enriched for and are known to be critical for lignin degradation [[68](#_ENREF_68)]. These enzymes do rely on oxygen of hydrogen peroxide which are toxic to many anaerobic organisms. This may have led to catalase being the most prevalent lignin degrading gene in the non-aerated conditions as catalase protects obligate anaerobes from oxidative stress [[5](#_ENREF_5), [70](#_ENREF_70)]. For this same reason, it is also plausible that peroxiredoxin during thermophilic non-aerated conditions may have served the same purpose [[71](#_ENREF_71)]. While anaerobic organisms have been shown to possess lignin degrading pathways, the enzymes were non-functional without oxygen [[63](#_ENREF_63)].

A heavily conserved gene amongst all our treatments was K01179 (endo-1,4-beta-D-glucanase) although Figure 2.5A showed significant difference to the non-aerated metagenome during T1. The conflicting activity measurements could have been due to cellulase inhibition from potential anaerobic conditions during saccharification [[72](#_ENREF_72)] . The difference in pH conditions may have produced differentially optimal conditions for the endo-1,4-beta-D-glucanase as different optimal pH’s have been reported [[73](#_ENREF_73), [74](#_ENREF_74)]. The other enzyme significantly overrepresented within the non-aerated metagenomes was the K05349 (beta-glucosidase). This is linked to some of the dominant bacteria within the thermophilic non-aerated community (OTU ID’s: 565764, 749791, 257863, 130040). The genus *Ruminoccoccus* is fairly well studied and known to be cellulolytic corroborates what was found as well as *Ethanoligenens* is suggested as having beta-glucosidase [[75](#_ENREF_75), [76](#_ENREF_76)]. This would be the first instance of an *Ethanoligenens* not only surviving at thermophilic conditions, but being a dominant community member.

Pectinases did not represent a major portion of the metagenome as can be seen from Figure 4. This is probably due to low initial pectin concentrations (previous characterization suggests around 7.55%) and the extensive soxhlet extraction process [[77](#_ENREF_77)].

While tomato pomace can be considered a heterogeneous feedstock full of multiple components such as seeds, skins, and pith, the majority of the community and deconstructive function can be ascribed to only seven phyla often dominated by only a couple OTU’s. Understanding how the metagenome is collectively acting will serve to provide us further understanding in how to deconstruct not only tomato pomace, but lignocellulosic feedstocks as a whole. While the PICRUSt approach does not align with typical metagenome mining for novel degradative enzymes, it facilitates further analysis into how suites of enzymes align together to create a degradative community. This is critical as all lignocellulosic substrates will have to be considered for a viable biofuels system and different synergistic enzyme combinations are constructed and evaluated for potential biofuels applications [[78](#_ENREF_78), [79](#_ENREF_79)]. There are even opportunities to identify niches where enriched organisms are also capable of producing biofuels themselves, such as *Ethanoligenens*, that may be taken advantage of when considering useful genes or organisms [[80](#_ENREF_80)]. Of even newer consideration is the creation of synthetic communities where only key community members are utilized to create symbiotic effects for biofuels production from complex substrates [[6](#_ENREF_6), [81](#_ENREF_81)]. Most synthetic communities exist as co-culturing but significant opportunities exist to expand into complex constructed microbial communities [[82](#_ENREF_82)].

## Conclusion

We enriched compost-derived microbial communities on water and ethanol extracted tomato pomace and showed changing community and genetic makeup. Aerated communities showed an enrichment of *Proteobacteria* and potential ligninases while non-aerated communities became enriched in *Firmicutes* and hemicellulases. Endoglucanase activities were not significantly different from each other and saw decreasing trends over time. This matched with the PICRUSt produced metagenome that showed no significant difference between treatments. Using PICRUSt on the 16S rRNA gene samples allowed us greater insight into the potential function of the community as well as allowed us to discern environmental relationships to gene prevalence. A collection of the most prevalent enzymes throughout the experiment were endo-1,4-beta-glucanase (K01179), alpha-fucosidase (K01206), alpha-N-arabinofuranosidase (K01209), and vanillate O-demethylase monooxygenase (K03386). Pectinases most likely did not play a significant role. Most potential ligninases in the non-aerated community may serve alternative roles in the protection from oxygen exposure. It may be useful in the future to further understand how to link the shape of the microbial community at hand to its environment. And to improve that understanding which OTU’s make up the majority of the function and how the OTU’s relate to specific elements of that environment will provide further insight into the development of synthetic microbial communities and subsequent understanding of plant-microbe interactions.

# Network meta-analysis method development and validation

## Abstract

## Introduction

Network analysis of microbial communities is allowing the opportunity to understand microbial community interactions that are otherwise infeasible to examine in the laboratory. The network analysis utilizes 16S rRNA gene sequencing to identify OTU’s and the infers relationships between the different OTU’s based on the compositional data acquired. Networks can be created using multiple open-source tools, such as MENA, SparCC, Spiec-Easi, and CoNet. Each of these tools can provide this sense of the interactions present between microbes [[39](#_ENREF_39), [83](#_ENREF_83), [84](#_ENREF_84)]. One consensus agreement that these network tools share is the need to prune spurious connections and that the true microbial network is “sparse” [[39](#_ENREF_39), [83](#_ENREF_83)]. Using these tools has started to give connections to commercially relevant crop diseases such as *Fusarium* wilt and provide ideas on bacteria and fungi that may suppress this pathogen [[85](#_ENREF_85)]. As microbial community information continues to be deposited in databases such as QIITA, NCBI, and IMG/M, integrating these studies to understand the microbial interactome will provide a unique ability to analyze microbial interactions on a much larger scale.

Integrating diverse environments will bring together many microbes and their representative sequences that are not typically observed together in the same environment. What this mean for the compositional data analysis is as you increase the number of sites you wish to include, sparsity will increase. Major problems with this compositional data will be sparsity due to sampling and lack of a true scale [[86](#_ENREF_86)]. This sparsity can be largely attributed to experiments and target niches which can create systematic sparsity (sparsity due to site sampling) and may result in applying filtering methods to construct relevant networks [[22](#_ENREF_22)]. Systematic sparsity can reduce the sensitivity of the detection methods to less than 10% by the time 50% or less of the OTU’s are shared between sites [[22](#_ENREF_22)]. To overcome this systematic sparsity, habitat filtering usually occurs to remove OTU’s that do not occur in “X” amount of the habitat where “X” can be an arbitrarily chosen number. Berry and Widder (2014) suggest filtering around 20% of co-occurrence, but this low-filtering can reduce the sensitivity of detection significantly. However, increasing the filtering above 60% dramatically decreases the specificity of detected edges. This is most likely attributed from the previous filtering of true interactions among minimally occurring OTU’s.

This has prompted many new efforts into improving how networks are constructed for microbiomes with such novel detection methods such as SparCC and Spiec-Easi as well as typical detection methods such as Pearson’s and Spearman’s which can identify linear and non-linear relationships respectively [[83](#_ENREF_83), [84](#_ENREF_84)]. These techniques are still error-prone given typical problems of sequencing data such as sparsity and low sampling [[21](#_ENREF_21)]. As many networks seek to uncover a set of core microorganisms for a given operation/environment, samples are usually acquired from many different sites/individuals and then combined and filtered by co-occurrence. This method inherently removes information by removing OTU’s regardless of their relative abundance by site/individual. So this often leaves the researcher with the decision of tossing data in an effort to extract high-quality information or utilizing the extra data but producing a set of interaction with lower specificity [[22](#_ENREF_22)]. Given a limitation of sequencing efforts, this can leave the researcher at the intersection of deciding how many sites/individuals should be sampled versus how many samples should be acquired from a(n) site/individual.

In the field of meta-analysis, combining data allows researchers to incorporate more data to strengthen conclusions about their research. This field has been a rising topic in sequencing data sets as more and more data are placed in publicly available repositories and may be an acceptable method in network analysis to replace filtering while avoiding decreased sensitivity. As we would like to develop a method for network meta-analysis that can learn from each set of newly input data it is important to consider both the strength of the correlations as well as the significance that the correlation exists. While methods to combine p-values are still an active area of research, some standard methods for combining independent studies are a generalized-Fisher’s and Stouffer’s weighted method [[87](#_ENREF_87), [88](#_ENREF_88)]. This is critical to identifying whether interactions continue to be significant across all sites/environments sampled.

In addition to combining p-values from each study, network analysis requires a decision to be made about correlations to decide the true strength of a connection. By combining correlation information, will help to understand the true strength of the interaction in the network. There are predominantly two-methods for combining correlations and that is averaged Fishers z transformations and averaging Pearson derived correlation values [[89](#_ENREF_89)]. Corey et al. (1998) showed that bias in the Fishers z transformation outperformed averaged correlations (r) in the case of fewer samples combined while averaged correlations expected “very slight advantage” as the number of samples goes to infinity. While these are not applicable for several of the currently network methods listed above, it provides a starting point and room for future research to improve this area.

While many set of interactions are possible, ranging from competition to mutualism, previous work by Weiss and Van Treuren et al. suggest that currently the best recovered interactions are commensal, competitive, and mutualism (although the recovered interactions varied by network technique) [[21](#_ENREF_21)]. While it is the goal to be able to uncover all ecologically relevant interactions in a microbial community, detection methods have not yet been developed that can unravel these connections from NGS data. In this paper, we examine the effects of various methods for combining correlations and p-values along with the effect of sample size and percentage of shared OTU’s on the ability to recover network edges using meta-analysis methods.

## Materials and Methods

### Simulated Microbial Community Construction

Engineered Networks were constructed using code adapted from Weiss and Van Treuren et al (2016) for commensal, competitive, and mutualistic interactions [[21](#_ENREF_21)]. The community data sets were generated from a log-normal distribution with X and Y and with zero-sparsity intra-site/environment. Sample sizes were kept small to emphasize that if the effect can be observed in simple low-OTU (20 – 200) count and low sample situations, more extreme situations would only be exacerbated further. For evaluation of combinatorial effects, one sample set with 20 “sites” was generated and split into equal smallest possible sample sites/environments (n=5) and compared to the full data being run through the same network detection pipeline.

To evaluate the systematic sparsity inter-site/environment, a 10 sample data set was split into two equal sample size “sites” and modified to simulate 25%, 50%, and 75% sparsity among OTU’s in the network. The true network across the three sparsity levels remained identical and the probability of a true connection existing between OTU’s found in both “sites” followed one minus the sparsity. OTU’s found in both “sites” were averaged using Fishers Z transformation and Fishers method for p-value combination

### Network Construction

Simulated microbial communities were then sent to SparCC for detection of interactions [[83](#_ENREF_83)]. SparCC relies on an iterative process to detect, exclude, and assign correlations. The detection process was iterated through 100 times and p-values were determined empirically through 100 permutations except where noted. The effect of exclusion and number of bootstraps on detection was examined using 100 and 500 iterations respectively.

### Correlation Combination

Two methods were used to evaluate the effect of combined correlation methods: average r’z and average r. The Fisher transformation converts the original correlation into a *z* using Equation 3.1 before being averaged and back-transformed using into the mean correlation value

Equation 3.1

Equation 3.2

Where:

r = original correlation value

z = transformed original correlation value

= mean correlation value

= mean of the transformed correlation values

### P-value Combination

While multiple methods exist, the Fishers and p-value averaging methods were compared for their applicability to network recovery

Equation 3.3

Where:

k = number of p-values to be combined

pi­ = ­p-values to be combined with an index of i

### Comparison of Results

Results were visualized with contour plots where the x- and y-axes are represented by p-values and correlation values respectively and the z-axis is informedness (also known as Youden’s J statistic or Youden’s index). Informedness is the mathematical difference between the true positive rate (TPR) and the false positive rate (FPR). The multidirectional approach allows the visualization of both filter typically applied to accepting an interaction between an OTU.

## Results

The informedness achieved by SparCC is considered the standard to which results will be compared.

### Network recovery

Examination of network recovery was tested against the original SparCC network tool and using SparCC in a piecewise manner and combining both the correlation coefficient (r) and the permutation derived p-value. In Figure 3.1, a comparison is made between Fisher’s method that transforms the correlation coefficient to a z value prior to average and back-transformation and that of averaging the correlation coefficient. The standard of SparCC (Figure 3.1 C) with all samples included showed that an Informedness greater than 0.949 could be achieved. Achieving such a high informedness could be achieved by using low p-values and a correlation coefficient less than 0.7. Going above 0.7 saw decreasing informedness from the loss of detection of true interactions. Fisher’s Z using 10 samples (Figure 3.1 A) shows that after combining two sets of five samples informedness can be recovered but over a smaller area and having to accept higher p-values. Averaging the correlation coefficient (Figure 3.1 B) showed a very narrow window in which true connections could be recovered and saw a much lower correlation coefficient threshold was needed. After combining all 20 samples (Figures 3.1 D & E), little difference existed between Fisher’s Z transformation and averaging the correlation coefficients. The averaging method show a slightly lower correlation coefficient that must be accepted but overall trends between the two methods remained nearly identical. While the two methods saw a much smaller swath in which p-values and correlation values could achieve as high informedness as the standard in Figure 3.1C, the averaging methods showed a markedly smaller dropoff (decreased slope on the contour plot) at lower and minorly higher correlation coefficient thresholds.

|  |  |  |  |
| --- | --- | --- | --- |
| (S) | Complete Data | Fisher’s Z | Average R |
| 10 |  | C:\Users\Josh\Google Drive\Experiments\Synthetic Data\iterative_test\ROC\ROC_AB.pngA | C:\Users\Josh\Google Drive\Experiments\Synthetic Data\iterative_test\ROC\ROC_AB_avgR.pngB |
| 20 | C:\Users\Josh\Google Drive\Experiments\Synthetic Data\iterative_test\ROC\10_All_ROC.pngC | C:\Users\Josh\Google Drive\Experiments\Synthetic Data\iterative_test\ROC\ROC_ABCD.pngD | C:\Users\Josh\Google Drive\Experiments\Synthetic Data\iterative_test\ROC\ROC_ABCD_avgR.pngE |

Figure .: Figures A-E are organized by number of samples included in the analysis (rows) and method applied (columns). Informedness was plotted showing areas of where filtering produces a more/less informed network. Fisher’s Z transformation is shown after combining 10 total samples (A) and 20 total samples (B). Average R is shown for 10 (B) and 20 (E) samples respectively. All are compared to running all 20 samples through SparCC (C).

Next, network recovery was examined through p-value combination using Fisher’s method (Figures 3.2 A and D) and through the average (Figures 3.2 B and E). The SparCC produced network that was constructed using all samples again is in Figure 3.2C. Correlations were combined using Fisher’s z transformation in all of Figure 3.2. After only combining 10 samples (Figures 3.2 A and B), the difference in methods show large deviations from each other. In Figure 3.2A, Fisher’s method shows a slight lower bound for the p-value filtering to still return a informedness of 0.949, but maintains the integrity of the p-value. The averaging of the p-value at 10 samples (Figure 3.2B) appears to be able to still achieve an informedness value of 0.949 but does so in a very awkward region of the graph. No longer are the true p-values maintained as significant, but rather the true interactions are being pushed into an elevated region (~ >0.2) of p-values. As 20 samples are averaged together, the integrity of the p-value from Fisher’s method (Figure 3.2D) is maintained and the upper threshold of p-value filtering for low correlations existing in the informedness region of >0.949 was reduced and is closer to that of Figure 3.2C. While in Figure 3.2E, the effect of averaging p-values shows a very wide range of acceptable p-value that produce an informedness >0.949, but lower p-values result in the premature filtering of true interactions. Another difference between the two methods is that the effect of filtering show’s a larger dropoff in informedness for p-value averaging (Figure 3.2E) while going to a decreased slope in Fisher’s method (Figure 3.2D).

|  |  |  |  |
| --- | --- | --- | --- |
|  |  | Fisher’s Method | Average P |
| 10 |  | A | B |
| 20 | C:\Users\Josh\Google Drive\Experiments\Synthetic Data\iterative_test\ROC\10_All_ROC.pngC | C:\Users\Josh\Google Drive\Experiments\Synthetic Data\iterative_test\ROC\ROC_ABCD.pngD | E |

Figure .: Figures A-E are organized by number of samples included in the analysis (rows) and method applied (columns). Informedness was plotted showing areas of where filtering produces a more/less informed network. Fisher’s method for p-value combination is shown after combining 10 total samples (A) and 20 total samples (B). Average P is shown for 10 (B) and 20 (E) samples respectively. All are compared to running all 20 samples through SparCC (C).

### Sparsity

This examined the feasibility of sample splitting across systematic sparsity (sparsity due to sampling site/environment) by comparing the result of a SparCC without splitting or co-occurrence filtering and by testing splitting and recombining using Fisher’s method of p-value combination and Fisher’s z transformation. Sparsity was discretely tested at 25%, 50%, and 75% sparsity. At 25% sparsity, systematic recombination (Figure 3.3B) shows a larger area in which maximum informedness is achieved but increased detection of false positives lower correlation thresholds. Both show largely decreased ability to achieve an informedness greater than 0.949. Moving towards increased sparsity, SparCC (Figure 3.3C) shows even further decreased ability distinguish in between false positives and true positives while splitting the data (Figure 3.3D) maintains a large area of informedness albeit only above 0.897. Splitting the data also requires a slightly higher correlation coefficient before it is stable against p-value choice. Among the systematic splitting (Figure 3.3D), there does appear to be a decreased slope for increasing p-values just as noted in Figures 3.1 and Figures 3.2. Going to 75% sparsity no systematic splitting (Figure 3.3E) results in significantly decreased ability to return a high informedness with a majority of the area returning informedness values less than 0.795. At 75% sparsity and systematic splitting of the data (Figure 3.3F) shows very little decrease in ability to return information with a maximum informedness less than 0.949 but greater than 0.897. A majority of the informedness values returned in Figure 3.3F were greater than 0.846. There was also a major decrease in the dropoff of informedness values returned at decreasing correlation values and increasing p-values (Figure 3.3F).

|  |  |  |
| --- | --- | --- |
|  | No Systematic Sample Splitting | Systematic Splitting and Fishers Recombination |
| 25% Sparsity | C:\Users\Josh\Google Drive\Experiments\Synthetic Data\Overlap_test\ROC\Over_25.pngA | C:\Users\Josh\Google Drive\Experiments\Synthetic Data\Overlap_test\Overlap_test_2\ROC\over_25.pngB |
| 50% Sparsity | C:\Users\Josh\Google Drive\Experiments\Synthetic Data\Overlap_test\ROC\Over_50.pngC | C:\Users\Josh\Google Drive\Experiments\Synthetic Data\Overlap_test\Overlap_test_2\ROC\over_50.pngD |
| 75% Sparsity | C:\Users\Josh\Google Drive\Experiments\Synthetic Data\Overlap_test\ROC\Over_75.pngE | C:\Users\Josh\Google Drive\Experiments\Synthetic Data\Overlap_test\Overlap_test_2\ROC\over_75.pngF |

Figure .: Figures A-F are organized by sparsity (rows) and method applied (columns). Informedness was plotted showing areas of where filtering produces a more/less informed network. Figures A, C, and E are output from SparCC without having been split prior. Figures B, D, and F have been split systematically to represent different environments and then were recombined using Fisher’s method for p-value combination and Fisher’s Z transformation

## Discussion

During network recovery, it was of interest to examine both whether or not information could be recovered through the use of combining p-values and correlation coefficients without the complexity of sparsity. This would allow the network tool (SparCC) to operate in most optimal conditions and would provide insight into the consequences of information that may be lost by combining correlation coefficients and p-values for OTU-OTU interactions that exist in multiple of the samples. Two basic methods, Fisher’s z transformation and averaging, were examined for the correlation coefficient and two methods, Fisher’s method and averaging, were investigated for combining p-values. This was meant to provide a rough guideline on the feasibility of such a technique but to leave room open to optimize in the future. During network recovery, averaging the correlation coefficients (Figure 3.1) demonstrated that not only was it feasible to combine the correlation coefficients, but that the method of combination is robust against user choice of correlation combination method. The major drawback of the averaging techniques was an increase in sensitivity to the choice of a user’s choice of p-value and correlation coefficient as a threshold. This is critical as networks of the microbiome are not known in advance and allowing flexibility in a user’s choice allows for robust analysis. However as Figures 3.1 D and E suggest, information can be recovered for high values of the correlation coefficient and lower p-values. The averaging of correlation coefficients does also tend to decrease the effect of poor p-value and correlation coefficient threshold values as the slopes on the contour plot are less severe. This may suggest that as more samples are combined, beyond 20, not only will the severity of a choice be made, but more of the false positives will be removed from the majority of the graph and informedness will increase.

As networks typically determine an interaction to be true based not only on correlation coefficient, but the significance of that interaction, it is vital to retain that information and therefore two methods were examined for their feasibility (Figure 3.2). Fisher’s method was better at maintaining the integrity of the p-value by decreasing the p-values associated with the true interactions while just averaging the p-value resulted in decreased detection of true positives at lower p-values. Both p-value combination methods saw an increase in the upper threshold for maintaining highest informedness at low correlation coefficients and this is likely due to the variations in the p-values for the true interactions. The power of the test for the correlation coefficient may be improved by moving beyond Fisher’s method [[87](#_ENREF_87)]. Focusing on p-value combination methods may be the best area to focus on future meta-analysis work as p-value threshold cutoffs appear more sensitive to recovering the largest area of informedness.

Splitting of data is shown here (Figure 3.3) to be beneficial to improving network interaction detection. SparCC appears capable of handling some deviation from zero-sparsity, but can benefit from meta-analysis for systematic sparsity. Weiss and Van Treuren et al. showed a similar decrease as sparsity increased [[21](#_ENREF_21)]. Our results show increased recovery because of the ecological interactions investigated in this paper targeted recovery of mutual interactions, but would likely extend to all ecological interactions. The method described in this paper will not likely improve the ability to recover more ecological interactions than the network interaction detection tool used because this method targets the removal of false positives more than it targets uncovering more true positives. One unique property this may open up is not just linking systematic sparsity (sampling due to site/environment) but random sparsity, a challenge of more likely caused by sequencing depth. This would be accomplished through the use of hierarchical clustering methods being used on sites and OTU’s and splitting samples by smallest available subset. This would effectively group random sparsity such that it could be treated as systematic sparsity.

## Conclusion

This paper demonstrated the feasibility of combining microbial community interaction to recover the underlying true network by using meta-analysis methods. There appears to be only a minor drawback in identifying the true network between recurring samples that is mostly mitigated after two smallest available samples are recombined. Methods surrounding the combination of p-values remains the method most available to improvement. Methods for combining correlation values does not appear to have a major effect on the network recovery. As sparsity increases, combining smallest available subsets of a dataset appear to outcompete the normal network interaction detection. The methods outlined here should be compatible with any network detection method although only SparCC was tested here.

# Meta-analysis of solarization communities using network analysis

## Abstract

## Introduction

The current norm in agriculture is to use synthetic chemicals for pest-management that can largely utilize methyl-bromide and chloropicrin [[90](#_ENREF_90)]. This can affect local air quality as well as global aspects [[91](#_ENREF_91), [92](#_ENREF_92)]. The continued use of fumigants allows the industry to be profitable where it is currently thought that the non-use would cause the U.S. farmers to be unable to compete internationally. Several promising techniques exist in various states of commercial availability and applicability such as anaerobic soil disinfestation (ASD) and solarization [[93](#_ENREF_93), [94](#_ENREF_94)]. Both of these aforementioned techniques are thought to modify the soil community and combat pathogens using microbial-produced control agents. One of the relevant secreted metabolites are volatile fatty acids (VFA) that are thought to have an effect on disease suppression in agricultural applications [[95](#_ENREF_95)]. Targeted within VFA’s tend to be formic, acetic, propionic, and n-butyric [[96](#_ENREF_96)].

Combining elements of both of these methods is biosolarization which incorporates organic elements into the soil (like ASD) and then uses solarization to heat the soil which then can benefit from added microbial heating and the secretion of microbial VFA’s to act as control agents [[94](#_ENREF_94)]. A critical element to the success of biosolarization is the microbial community present during the narrow window of heating. It is known that substrate and depth play a role in the microbial community present but only one publication has begun to examine the interactions

of the microbial community responsible for biosolarization [[10](#_ENREF_10)](achmon future publication).

Each experiment sought to understand the effect of substrate on the shifting microbial community and the subsequent weed-inactivation but substrates can be extremely seasonal and the ability to find the substrate of interest may impact the ability to adopt such an agricultural practice [[49](#_ENREF_49)]. Fernandez-Bayo et al (2017) did study a more seasonally robust substrate but microbial communities from digestate can also vary widely [[8](#_ENREF_8), [97](#_ENREF_97), [98](#_ENREF_98)]. As the environment can largely impact the community, the biosolarization environment may provide a more extensive selective pressure than the community itself [[46](#_ENREF_46)]. While the environment can be a tremendous selective pressure, microbial interactions account for another significant pressure when shaping the community and can even help shape their environment [[46](#_ENREF_46)]. In solarization, the community is responsible for the deconstruction of the lignocellulosic substrate and subsequent production of VFA’s. Currently it is not known who within the community is responsible for VFA production, lignocellulosic deconstruction, and whether they are the same set of microorganisms.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Experiment | Substrate | Samples | Year | Reference |
| Simmons et al. | Wheat Bran | 42 | 2012 | [[10](#_ENREF_10)] |
| Achmon et al. | Tomato Pomace | 75 | 2015 |  |
| Fernandez-Bayo et al. | Thermophilic/Mesophilic Digestate | 65 | 2015 |  |

Table .: Primary author, tested substrate, number of samples utilized from experiment, and year the experiment was run

At the heart of the community are potentially millions of interactions between every OTU that can result in changes to the shape of the community. To examine microbial interactions, network analysis has become an excellent tool for evaluating positive and negative interactions of different organisms [[99](#_ENREF_99)]. Understanding these interactions can provide insight into symbiotic interactions and identify sub-communities that may be filling a working niche [[21](#_ENREF_21), [26](#_ENREF_26)]. Attempting to identify true interactions among the millions of possible interactions can be difficult with programs like SparCC and Spiec-Easi designed to identify and remove spurious interactions [[83](#_ENREF_83), [84](#_ENREF_84)]. These network programs work on the fundamental principle that most OTU’s are not interaction with one another and this is vital to returning quality information that can be analyzed.

As interactions are identified, localized clustering tends to occur resulting in these important sub-communities. To start to decipher the community as a whole, it is important to identify these sub-communities using community detection methods using robust methods [[19](#_ENREF_19), [42-44](#_ENREF_42)]. While many methods are available and present pros/cons associated with different network structures, initial exploration of sub-communities can explore detection algorithms using modularity optimization to identify larger sub-communities [[100](#_ENREF_100)]. Other studies have started to provide heuristics for available clustering algorithms and when to use them based on size and mixing parameter [[20](#_ENREF_20)]. Once sub-communities have been identified, characterizing which sub-community is responsible for different functions within the overall solarization community or how they are impacted by different environments can provide significant insights into community dynamics [[101](#_ENREF_101)]. With sub-communities in solarization can allow further investigation into the environments they are found in future efforts that may allow manipulation of the sub-communities such that solarization time can be decreased and/or the effectiveness of pest management can be increased.

Although microbial community data exists for multiple biosolarization experiments (Table 4.1), there currently exists no continuity in the information surrounding the microbial community.

By combining the information from the microbial community, the available number of samples goes from 42 to 182. As the number of samples increases, the quality of information in network analysis increases [[22](#_ENREF_22)]. By conducting the analysis across a wider array of substrates and samples, the overall analysis becomes can have applicability across a greater set of soils and substrates that may eventually be incorporated.

The goal of this research was to analyze multiple solarization microbial communities to identify to true underlying interactome present in solarization. Microbial 16S rRNA gene data sets were split into smallest possible subsets to be run through SparCC’s correlation detection. Each subset of data was later recombined and analyzed for network structure, sub-community structure, and connected with the VFA production from the original experiments. This data can inform future target microbes that are key to solarization as a whole via identifying sub-communities responsible for different characteristic lignocellulosic deconstruction pathways and VFA production.

## Methods

### DNA-sequencing and Analysis

DNA was sequenced as described in respective studies and FASTQ sequences were retrieved. FASTQ sequences were then aligned against the Greengenes Database (version 13.8) using QIIME (version 1.9.1) and default settings and 97% sequence identity [[29](#_ENREF_29), [57](#_ENREF_57)].

### Correlations

Samples from each dataset were grouped into smallest available groupings of sample size greater or equal to five. OTU’s in each subset of data were filtered by having a mean of 10 reads per sample (approximately 0.05% of relative abundance) to reduce sample complexity [[102](#_ENREF_102)]. Each subset was then sent to SparCC for 100 iterations [[83](#_ENREF_83)]. 100 bootstraps were generated to determine an empirical p-value. All p-values and correlation values were then stored in a SQLite database (version 3.8.3) for use later.

### Network Construction and Analysis

Networks were built by pooling p-values and correlations from the database that occurred in at least two out of the 31 subsets. P-values were combined using the Fisher method [[87](#_ENREF_87)]. Correlation values were combined using Fisher’s Z transformation [[89](#_ENREF_89)]. A link was created when both criterion of a combined p-value ≤ 0.1 and a combined correlation value ≥ 0.6 were met.

The Multilevel (Louvain) community detection method by Blondel et al. was used to identify sub communities [[20](#_ENREF_20), [43](#_ENREF_43)]. Network connectivity of nodes was evaluated using a Zi-Pi plot (see ref Olesen et al. for simplification and ref Guimera and Amaral for original concept) [[103](#_ENREF_103), [104](#_ENREF_104)].

### Predicted Metagenomics

PICRUSt (version 1.0.0 dev) was used to identify predicted metagenomes for all OTU’s included in the network. Pathways were then consolidated into level 1-3 as part of PICRUSt *categorize\_by\_function.py*. Metagenomic samples were evaluated for deconstructive abilities and as such samples were filtered for respective KEGG orthologs as described previously [[38](#_ENREF_38)](tp decon).

A pseudo-count count was generated with each OTU in the network containing 1000 counts prior to metagenome prediction. After metagenomes were predicted successfully, the output count table was normalized by total cluster OTU’s and by subtracting the mean and dividing by the standard deviation for each gene in the table. Thus a value of 1.0 would indicate that the module has an average gene count per OTU that is 1.0 standard deviation above the average amongst all modules. Additionally, all KEGG orthologs identified near as catalyzing a reaction to a potential VFA were further analyzed.

|  |  |
| --- | --- |
| Volatile Fatty Acid | KEGG Ortholog |
| Acetate | * K00128 * K00138 * K00149 * K00467 * K00925 * K01026 * K01067 * K01512 * K01895 * K01905 * K14085 * K18118 |
| Butyric | * 00929 * 01034 * 01305 * 19709 |
| Lactate | * 00016 * 00101 * 07248 * 19266 |
| Propionate | * 00925 * 00932 * 01026 * 01895 * 01905 * 01908 * 19697 |

### Statistical Analysis

## Results

### Network Structure

The database resulted in 31 subsets of data and a total of 182 samples. In the database there are XX number of unique interactions and YY number of OTU’s. Upon construction of the network there were 1683 OTU’s and 16699 edges. The resulting network was considered small-world with an S∆ of approximately 26.5 [[105](#_ENREF_105)]. The degree distribution of the solarization interactome was related to a power law (Appendix X) and the average path length was 2.19. The network had one island with two OTU’s.

### Network Characterization

The constructed network consisted of 27 phyla and the top 10, visualized in Figure 4.1, made up 97.3% of the total distribution. The biggest contributions came from *Proteobacteria* and *Firmicutes*.

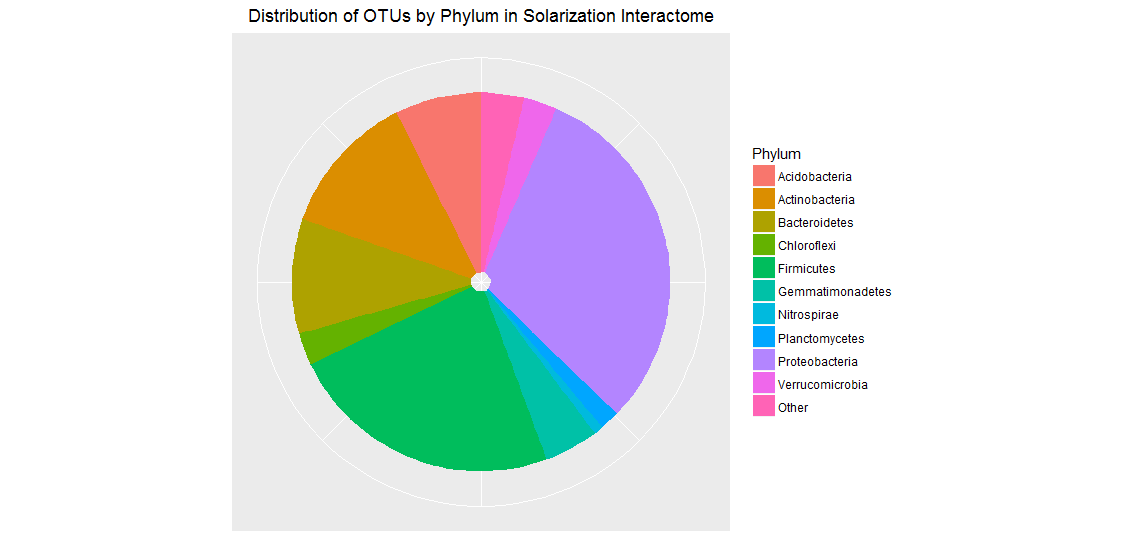


Figure .: Distribution of OTU's in solarization interactome (p-value ≤ 0.1; correlation ≥ 0.6)

The network was visualized with Gephi (Version 0.9.1) in Figure 4.2. Louvain community detection revealed 10 unique clusters. The cluster populations ranged from 316 (Cluster 3) in the largest cluster to 2 (Cluster 5) at the smallest. Cluster 5 was the island referred to in the previous section. The smallest cluster within the connected network was 62 OTU’s.

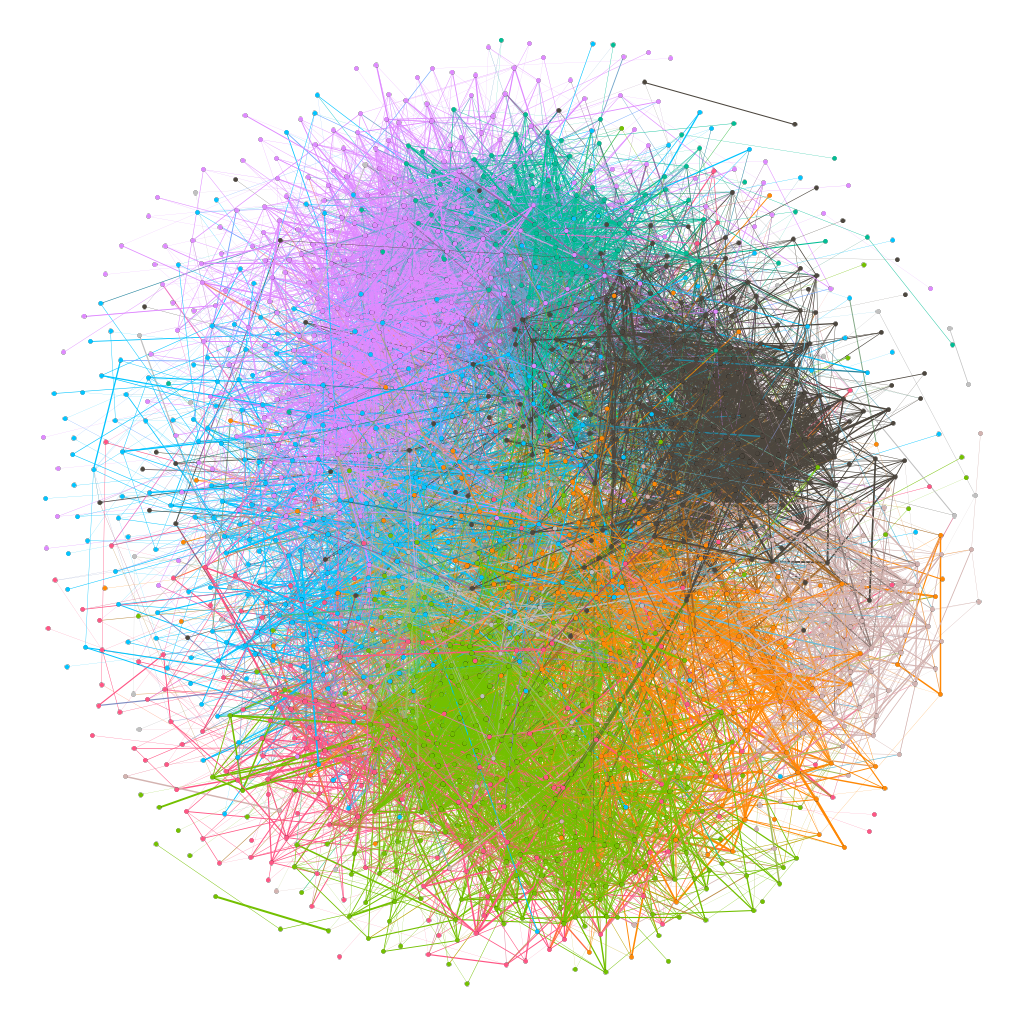


Figure .: Solarization network (Gephi v 0.9.1) with modules detected by Louvain method shown as different colors

The phylogenetic distributions of each cluster saw deviations from the entire network distribution of phylum (Appendix ). The number of members in the cluster did not appear to the influence the phylogenetic distribution of the clusters. Clusters were largely dominated by *Firmicutes* and *Proteobacteria*, similar to the whole network but saw deviations such as *Acidobacteria* in Cluster 4 and *Actinobacteria* in Cluster 9.

### Solarization Results

To examine the distribution of clusters by treatment and depth, a heatmap was generated (Figure 4.3) show relative abundance of each cluster. Cluster 6 was the most abundant across all depth and treatments but showed some narrow areas of increased density (tomato pomace with compost or wheat bran with compost in both the middle and lower soil layers). Cluster 8 while showing decreased relative abundance in all depth associated with tomato pomace inoculated with compost, showed enrichment when inoculated with mesophilic digestate. Cluster 2 was most represented in a soil/compost mixture but showed decreased enrichment in soils amended with wheat bran inoculated with compost, tomato pomace, and with tomato pomace inoculated with compost. Cluster 10 Show increased abundance in the top layer of the soil and additional in the medial layer for the wheat bran inoculated with compost. Cluster 3 had the greatest abundance occur in compost/soil mixtures and soil by itself with any further amendment reducing its relative abundance. The remaining clusters were present in decreased quantities but showed narrow preferences for treatment and depth.

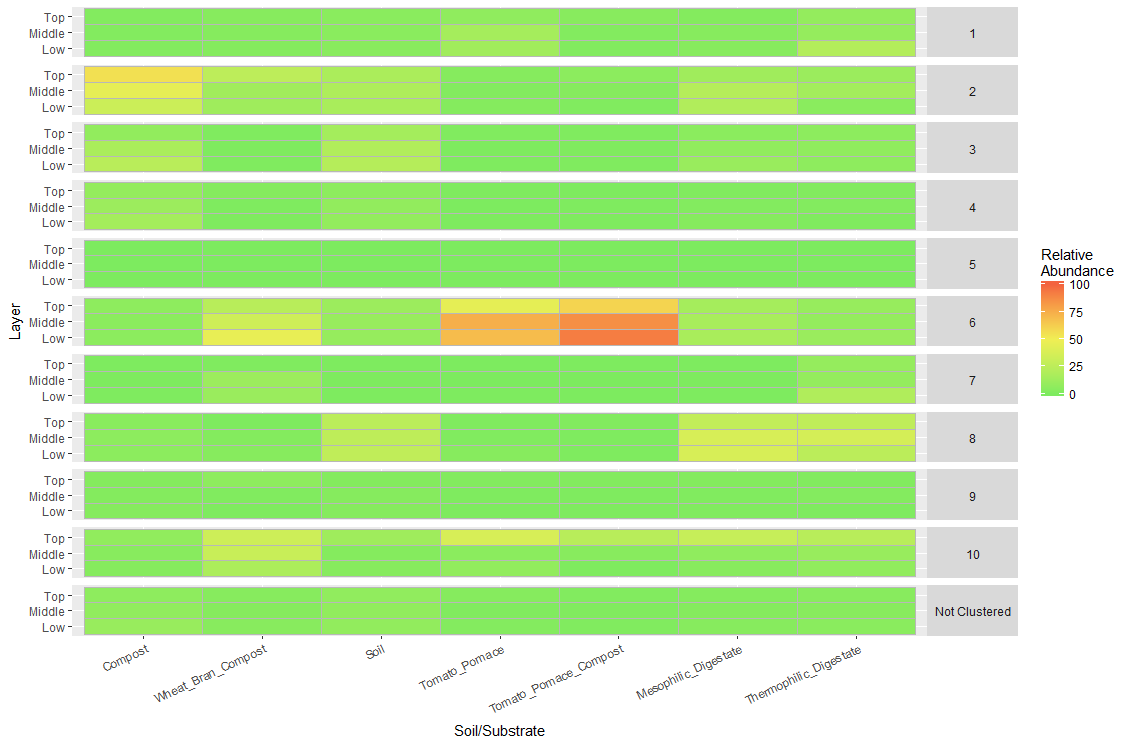


Figure .: Heatmap of clusters present by treatment and depth in the studies included in the construction of the interactome.

Examining the correlation of clusters to VFA production (Figure 4.4) revealed unique clusters that are consistent with VFA production. Cluster 6 and 7 was correlated to VFA production across two of the experiments, although cluster 7 was negatively correlated to VFA production in Achmon et al. Cluster 1 in Fernandez-Bayo et al. was positively correlated to VFA production, but showed no other positive correlation and was negatively correlated in Achmon et al. Cluster 10 showed no significant correlation across all experiments. All remaining clusters were consistently negatively correlated to VFA production. Positive isobutyric correlations did not always correspond to a positive total VFA production but acetic acid correlation remained consistent with total VFAs.

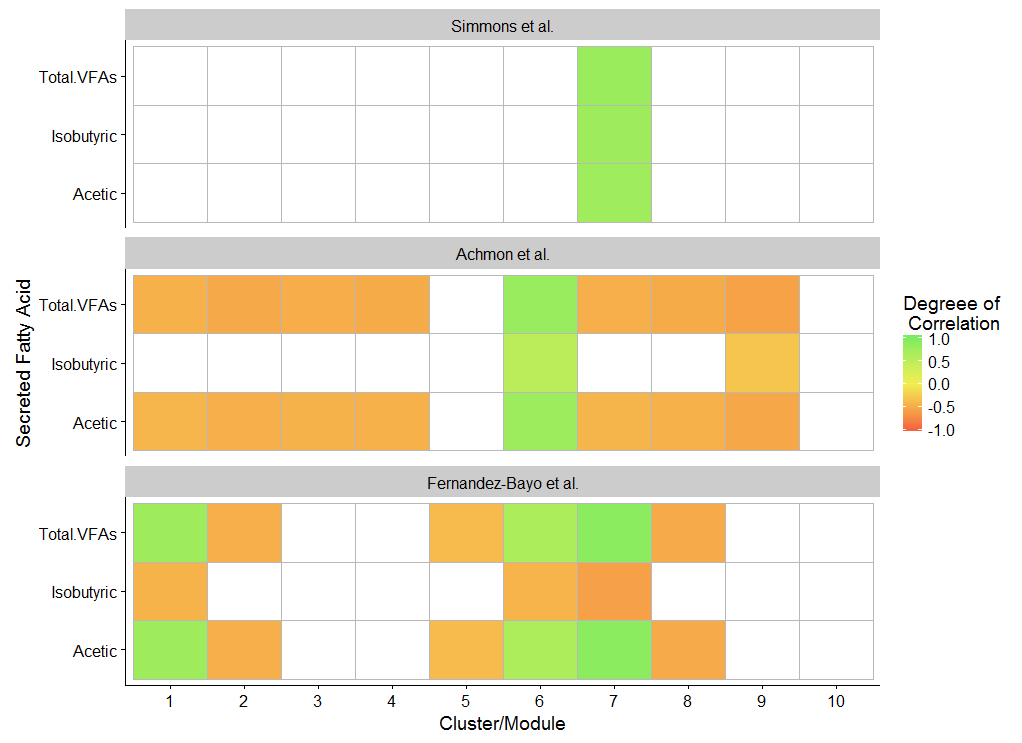


Figure .: Correlations of clusters to both individual and total volatile fatty acid production

To examine the network connectivity of the OTU’s within the network, a Zi-Pi plot was created. The Zi-Pi plot (Figure 4.5) allows the identification of organisms critical to the structure of the network. The Zi-Pi plot is divided into four zones; zone1: low intra-module connectivity, low inter-module connectivity; zone 2 (Connectors; ●): low intra-module connectivity, high inter-module connectivity; zone 3 (Hubs; ●): high intra-module connectivity, low inter-module connectivity; zone 4 (Cornerstone; ●): high intra-module connectivity, high inter-module connectivity. Identified within the Zi-Pi plot was 1149 bacteria classified as Connectors, 43 as Hubs, and 13 as Cornerstone.

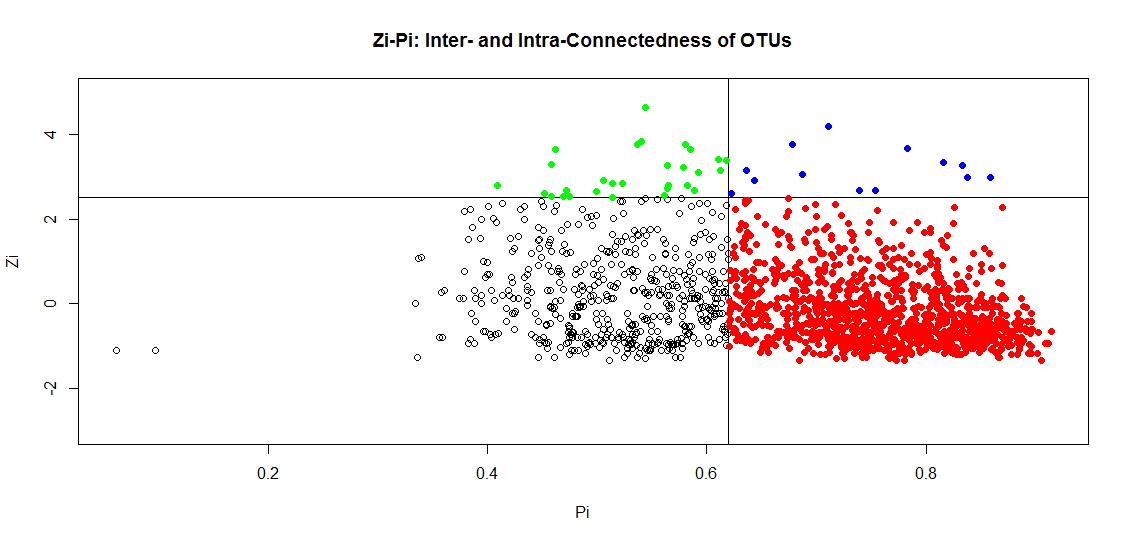


Figure .: Zi-Pi plot displaying the connectedness of OTU's. Green circles (●) represent hubs, red circles (●) represent connectors, and blue circles (●) represent cornerstone OTU’s.

As Cornerstone OTU’s are widely connected to multiple modules as well as deeply connected to their own module, they are thought to be a critical founding element for the construction of these communities. Relative abundance of the OTUs marked as cornerstone to be present in a majority of samples, but in low abundance (<1.1%) of their average in the total community (Appendix ).

In Figure 4.5A and Figure 4.5B the heatmap was produced for the top five Hubs and 10 Connectors (as found by average relative abundance across all treatments) Several OTU’s could be identified to specific treatments, inoculum, and depth. OTU 594040 (*Flavisolibacter*) showed a preference for all depths in soil and digestate inoculated samples but decreased when compost or tomato pomace was present. OTU 940737 (*Aquincola tertiaricarbonis*) showed a similar trend but also showed a slight preference for higher depths. Several hubs showed a preference for tomato pomace: OTU 543607 (*Clostridiaceae*) and OTU 679939 (*Acetobacter*). Connectors showed a wide array of preferences across treatments and depth. An OTU like 1051517 (*Bacillus*) showed a preference for tomato pomace amended samples and mesophilic digestate, but were near ubiquitously present across all samples in a relatively large amount for a single OTU.

|  |
| --- |
| A |
| B |

Figure .: Heatmap of the top 5 hubs (A) and top 10 connectors (B) by relative abundance in both depth and sample.

### Metagenome

Using PICRUSt to reconstruct metagenomes revealed similarities within the clusters among the deconstruction related genes (Figure 4.7A). Certain genes were more enriched than others across the entire solarization interactome. Critical genes appear to be K01179 (endo-glucanase), K01187 (alpha-glucosidase), K05349 (beta-glucosidase), K05350 (beta-glucosidase), K01206 (alpha-L-fucosidase), K01209 (alpha-N-arabinofuranosidase), and K00104 (glycolate oxidase). Although the most consistent, cluster 6 did not appear to have any unique deconstruction related gene. Cluster 8, most-enriched in the digestates showed a slight preference for K00428 (cytochrome c peroxidase) and K00432 (glutathione peroxidase). Unique to cluster 6 was an enrichment of phosphotransferase genes (PTS) associated with transporting cellobiose across the membrane (appendix). Looking at the fermentative genes product (Figure 4.7B) again saw no distinct difference between the clusters, but did see a unique preference for genes (K00128, aldehyde dehydrogenase; K01895, acetyl-CoA synthetase) within the fermentative pathways. Acetyl-CoA synthetase is also present in the acetic acid pathway as the gene is responsible for many reactions involving CoA.

|  |
| --- |
| A |
| B |

Figure .: Heatmap of relative abundance of genes within each cluster using 1000 counts per OTU to generate total counts for (A) deconstructive genes and (B) fermentative genes.

Just as with clusters, correlations between genes were examined for their relationship to VFA production (Figure 4.8). Most genes were consistent in the direction of their correlation to acetic acid and total VFA production except K01191 (beta-mannosidase). Beta-mannosidase only saw positive correlation within the study by Fernandez-Bayo et al. Cellulose-degrading enzymes except K01179 (endo-1,4-beta-D-glucanase) saw either negative correlations or none at all. Hemicellulase genes were inconsistent among which genes would correlate to VFA production, but were consistent across all experiments. Lignin degradation was largely negatively correlated except for K05909 (laccase) and K11065 (thiol peroxidase). Laccase was not detected in Fernandez-Bayo et al.

|  |
| --- |
| A |
| B |

Figure .: Correlation of (A) deconstructive and (B) fermentative genes to VFA production

## Discussion

Despite the network being a very large number of OTU’s, we have shown the small-worldedness of such soil-based communities. The small-worldedness is critical to understanding the structure of the network and how information flows. This suggests that certain OTU’s remain clustered together but there exists very short path lengths between all OTU’s and one another which can facilitate the rapid passing of information between nodes (or in our case it is likely metabolites/DNA/etc between OTU’s) [[106](#_ENREF_106)]. The clustering of OTU’s was elucidated as sub-communities using community detection resulting in 10 clusters. Multiple community detections remain available, and have been noted as being able to detect very small clusters but the resulting quantity of clusters proves difficult to extracting critical information [[19](#_ENREF_19)]. It also seems likely that while smaller communities exist, they are part of a larger hierarchical function much like metabolism and this investigation can be pursued upon further examination of individual clusters. Localized enrichment of clusters appeared in different soil depths and treatments, but could be noted across treatments suggesting environment may play a role in driving these synergistic interactions [[46](#_ENREF_46)]. These microbial niches are likely to be the result of functional requirements of the microbial community [[107](#_ENREF_107)].

The 10 clusters identified included on island of disconnected Actinobacteria. Neither of the two OTU’s included in this cluster (cluster 5) represented a large fraction of any community and were deemed largely as potential background and not of significance to bio-solarization. Clusters were typically dominated by Proteobacteria or Firmicutes. Cluster 4 did contain a significant proportion of Acidobacteria which saw the largest relative abundance in soil only (7.7%) and soil inoculated with compost (14.8%) suggesting that it may have been part of the initial soil and compost as it represents approximately 30% of the compost inoculum (tp decon) but is not able to outcompete when extra organic matter is added. Clusters 2, 3, 7, and 8 contain a majority of Proteobacteria which correspond to community majorities in soil only, soil inoculated with compost, and both digestates. The prevalence of Proteobacteria in soils is a common trend that has also been associated with the “active layer” in previous permafrost studies [[108-110](#_ENREF_108)]. Coolen et al. suggests that this Phylum may thrive due to its ability to fix nitrogen and may not be a compete for local resources [[110](#_ENREF_110)]. Firmicutes was identified in the permafrost soils but typically at the deeper samples and Coolen et al. suggested that it may be its ability to thrive at colder temperatures and dessication. Although not thoroughly addressed in that study, potential oxygen deprivation may be a factor at depths as Firmicutes appears to thrive in tomato pomace deconstruction during unaerated conditions (tp decon). This may provide insight into clusters 1, 6, and 10 which contain a majority of Firmicutes. Although not typically seen in soils, regardless of organic content, Firmicutes makes up a majority of the soils amended with organic matter [[109](#_ENREF_109)]. This is most similar to soil disinfestation conditions (BSD) where Firmicutes made up 59.4% of the BSD community [[111](#_ENREF_111)]. Mowlick et al. also noted that upon removing the tarp typical in BSD and post-harvest, that conditions again were enriched towards Proteobacteria (45.6%). It is also noted that anaerobic bacteria were identified within the bio-solarization community, such as *C. butyricum* (cluster 6; OTU’s: 238205, 541328), *C. tyrobutyricum* (cluster 6; OTU: 16144), and *C. thermosuccinogenes* (cluster 10; OTU: 53642), suggesting again that oxygen deprivation was occurring in bio-solarization and that subsequently Firmicutes can be a dominant Phylum.

Most clusters did not represent more than 50% of any community except for cluster 6 that saw enrichment upwards of 90% within soil amended with tomato pomace and inoculated with compost. While cluster 6 saw enrichment in soil amended with tomato pomace alone, the addition of compost as inoculum had an interactive effect with tomato pomace that could not be explained by additive measures. Similarly, with soil amended with wheat bran inoculated with compost, cluster 6 was enriched. This would suggest that although some of cluster 6 can be enriched directly from the soil, adding organics and compost further increase the virility of this cluster within the environment. Cluster 6 also saw an increase in relative abundance as the depth of the soil increased suggesting a more temperate-loving cluster. In cluster 8, digestate appeared to invigorate OTU’s already present in the soil and that the digestate community did not take a large foothold in the solarization environment. Cluster 2 was mostly identified among samples of soil inoculated with compost but with no organic amendment. There also appears to be a gradient effect as the soil layer gets deeper suggesting cluster 2 is mostly thermophilic just like a composting environment would be. Cluster 10 appeared largely in the top layer of the soil across most treatments suggesting again a thermophilic relationship although 10 was largely diminished in soil only and soil inoculated with compost. This would suggest a cluster with a preference for thermophilic conditions and the need for organics, but the origin of the cluster is undetectable. Clusters 6 and 10 appear to be two critical halves of solarization with cluster 6 occupying more mesophilic temperatures and cluster 10 in the thermophilic regions. While cluster 6 is largely correlated to VFA production across all experiments, cluster 10 uniquely did not positively or negatively correlate with VFA production.

Network interactions can give insight into potential keystone microbes that are significant to that environment and may provide OTUs that play a critical role in maintaining the stability of the community or passing information [[26](#_ENREF_26)]. Generally, there existed a larger degree of “Connectors” than is seen in typical network studies [[26](#_ENREF_26), [85](#_ENREF_85)]. This is likely caused by low-filtering on the relative abundance with most OTU’s existing at less than 1% of the total community. This also is partially caused by deep sequencing of the 16S rRNA gene (>100,000 counts per sample as witnessed but will likely exist at lower counts) as low count OTUs would not be sequenced (analogous to rarefaction) but this has its own pitfalls [[112](#_ENREF_112)]. Although not tested at different sequencing depths, the analysis was repeated a second with less filtering on correlation strength and significance associated with the p-value and a majority of the same OTUs were identified and cluster characteristics appeared similar suggesting that this method may be robust to deviations from number of included OTU’s. More OTUs in the network also reduced the percentage of OTUs identified as keystone again suggesting that the method is robust against parameter selection. What is still indiscernible is whether or not some of these “keystone” organisms are critical or omnipresent but inactive due to the heterogeneity present in the soil. As ecological significance of these OTUs can’t be determined until post-analysis, erring on the side of caution and keeping the OTU’s rather than omitting through methods such as rarefying is recommended [[113](#_ENREF_113)]. Although this may be a unique opinion as others suggest omitting rare OTU’s prior to downstream analysis [[114](#_ENREF_114)].

Among the keystone OTU’s, to deepen the relationship between clusters 6 and 10, several interactions exist within the network. OTU’s 91980 (thermoamylovorans; cluster 6) is connected to both OTU 578347 (plannococceae; cluster 10) and OTU 322136 (thermoamylovorans; cluster 6) and represent a significant portion of the total relative abundance in their respective samples. Thermoamylovorans may have significance related to highly deconstructive communities although its individual role was not elucidated [[115-117](#_ENREF_115)]. The persistence of thermoamylovorans in the top layer coincides with a temperature increase and pH increase which both favor the growth of this organism [[118](#_ENREF_118)]. Just as with cluster 6, thermoamylovorans has the ability to grow on cellobiose and produce acetate as the primary fermentation product [[116](#_ENREF_116)]. While plannococcaceae has been identified in many environments, including those of lignocellulose degradation, limited information exists based on the limited resolution of taxonomy [[119](#_ENREF_119)]. A major hub was OTU 238205 (*C. butyricum*) and its presence indicates small pockets of strict anaerobes. As the depth of the soil increased, so did the relative abundance of this microbe indicating that at increasing depth during bio-solarization there will be an increasing oxygen deficiency. Interesting to note is that the main fermentation product of *C. butyricum* is butyric acid, but butyric acid was only ever detected in small quantities suggesting there exists a butyric acid oxidizing bacteria present. Typically, the family associated with butyic acid oxidation is *Syntrophomonadaceae* but none were detected across all experiments. What was identified was the family *Peptococcaceae* that contains genus such as *Pelotomaculum and Desulfotomaculum*. This highlights the potentially valuable role that cornerstone OTU 898095 (*Desulfotomaculum*) may play as *Desulfotomaculum* have been identified as being able to oxidize long chain fatty acids along with short chain fatty acids such as butyric acid [[120](#_ENREF_120)]. Although OTU 898095 is present in low relative abundance it may play a disproportionate role in the oxidation of fatty acids. Also identified within *Peptococcaceae* was *Desulfosporosinus meridiei* that can consume butyric acid as well may be responsible for the turnover of sulfur compounds [[121](#_ENREF_121)]. As Hausmann et al. noted, *D. meridiei* was present only in low abundance and yet was still a critical element in the sulfur reduction and management of methane production [[121](#_ENREF_121)]. Although the rare OTUs typically are thought to be background or a “seed bank”, Hausmann et al. was able to show activity among the rare OTUs and the cornerstone OTUs identified here may also fall into that role. It also suggests that these low-abundance OTUs may be responsible for the oxidation of butyric acid, among potentially other items, during bio-solarization and that sulfur may play a critical role in VFA cycling.

While roles yet unidentified may affect rare OTUs, many macroscopic properties exist within the metagenome such as deconstruction and fermentative pathways. While clusters did not exhibit extremely unique sets of deconstructive genes, not all deconstructive genes were implored to deconstruct wheat bran, tomato pomace, and the digestates. This may indicate there is a core set of genes associated with the rapid deconstruction of lignocellulosic feedstocks during solarization. The most common cellulose degrading enzyme was endo-1,4-beta-D-glucanase which has been identified previously as the most prevalent enzymatic activity during tomato pomace deconstruction and may play a critical role among all lignocellulosic substrates [[122](#_ENREF_122), [123](#_ENREF_123)]. In contrast, KO1225 (exoglucanase) did not represent an appreciable portion of the clusters metagenome. Endoglucanase was accompanied by beta-glucosidase (K05349 and K05350) to break down the dimers into metabolizable glucose. Clusters 6 and 10, containing the cellobiose transporter, also contained a higher level of K01222 (6-phospho-beta-glucosidase) for degrading the phosphorylated cellobiose within their clusters metagenome. The cellobiose transport system may provide a competitive advantage for the clusters 6 and 10 to become established as low concentrations of cellobiose can be rapidly transported intracellular [[124](#_ENREF_124)]. A common but likely not critical enzyme across all the clusters was alpha-glucosidase (K01187) although simple starches may be present during the initial few days.

Hemicellulase degrading enzymes were minimalistic across all cluster with items such as endo-1,4-beta-glucanase, alpha-galactosidase, beta-galactosidase, and carboxylesterase. In a previous study, alpha-L-fucosidase and alpha-N-arabinofuranosidase were heavily enriched in the anaerobic environment regardless of temperature. This would suggest these enzymes will be robust against the diurnal cycling associated with bio-solarization. Cluster 6 showed a decreased relative abundance of hemicellulases as compared to other clusters, but showed an increased evenness in the availability of hemicellulase genes. The genes for alpha-mannosidase (K01191) appears enriched only in cluster 6 and may be necessary to unlock more available cellulose within the anaerobic environment [[119](#_ENREF_119)]. This is reinforced by Figure 4.8A that suggests these hemicellulose-degrading genes are correlated with VFA production.

As it pertains to lignin-degrading genes within the predicted metagenome, most genes present within the clusters were negatively correlated with VFA production. The most common gene amongst the clusters for lignin degradation, K00104 (non-heme chloroperoxidase), may play a dual role for bio-solarization as Bengston et al. suggests this as a key enzyme for production of antimicrobial agents [[125](#_ENREF_125)]. Although VFAs have been identified as the major element responsible for pathogen control, this may represent a new avenue is potential mechanisms that are involved in pathogen control strategies such as bio-solarization and anaerobic soil disinfestation. The one lignin gene associated with VFA production, K05909 (laccase), can not only degrade lignin, but lignin degradation products [[126](#_ENREF_126)]. The coordination of laccase with peroxidases and oxygen among the radical species produced during the lignin degradation process are not typical of conducive to an anaerobic environment, but Lucey and Leadbetter suggest that anaerobic organisms may use these oxygen dependent enzymes both as a way of detoxifying the surrounding phenolic compounds and consuming the toxic, for strict anaerobes, oxygen [[63](#_ENREF_63)].

Examining the fermentative pathways among the cluster show consistent preference for three main enzymes: K00128 (aldehyde dehydrogenase), K00925 (acetate kinase), and K01895 (acetyl-CoA synthetase). Aldehyde dehydrogenase and acetyl-CoA synthetase show a mixed set of positive and negative correlations to VFA production suggesting that they may be inconsistent in terms of the dominant pathway during fermentation. While aldehyde dehydrogenase can lead to acetic acid, it can also convert acetate into acetaldehyde and subsequently ethanol. Although a previous study examines the conversion of lactic acid into acetic acid and 1,2 propanediol and they observed the preferred direction of aldehyde dehydrogenase was to produce trace amounts of ethanol [[127](#_ENREF_127)]. This supports our observation of a negative correlation to VFA production. The dominant pathway in bio-solarization for acetic acid production appears to be through acetate kinase; both a dominant enzyme within the clusters and highly correlated to VFA production as well as acetic acid. This pathway would allow the generation of extra ATP at the expense of producing a more acidic environment which has been reported as increasing the growth rates in *Lactobacillus helveticus* ATCC 15807 at pH 4.5 [[128](#_ENREF_128)]. Chidthaisong and Conrad showed that chloroform can inhibit acetate kinase despite acetate production increasing in their samples [[129](#_ENREF_129)]. This supports the idea that the non-heme chloroperoxidase is producing chloroform and would therefore inhibit the dominant acetic acid pathway utilizing acetate kinase although the heterogeneity in the soil may provide ample distance between the chloroform and the acetate producing organisms.

## Conclusion

This paper provides the first in depth look at the bio-solarization microbial community through meta-analysis, networks, and predicted metagenomics. Network analysis produced 10 distinct clusters that each filled a unique niche within the bio-solarization environment. Two distinct clusters were expected to have bio-solarization relevance with respect to VFA production. VFA production during bio-solarization appears to be a highly regulated process that involves sulfate reduction, lignin decomposition, and chloroform production. Better understanding these unique mechanisms and the organisms responsible for them can potentially increase the viability of bio-solarization as a commercially relevant tool in agricultural food production.

# The ability of artificial neural networks to predict microbial communities using the interactome and the current limitations

## Abstract

## Introduction

Prediction of the microbial community represents a valuable tool that can create advancements in health, agriculture, ecology , and more [[46](#_ENREF_46)]. Current prediction methods rely on interaction based upon phylum, construction of a Bayesian directed acyclic graph, and a machine learning tool (Eureqa; [www.nutonian.com](http://www.nutonian.com)) that establishes complex equations between nodes in the graph [[46](#_ENREF_46)]. This has been applied to acid mine drainage with various degrees of success depending on the taxonomic level to which information was binned [[130](#_ENREF_130)]. One of the major issues with prediction using taxonomic binning is that taxa can be observed across environments and this forces complex equations to describe the dynamics and errors in the prediction. The may be improved by understanding how specific microbes work together and using that extra information to our advantage.

To understand how microbes work together, network analysis of microbial communities aims to detect microbial interactions from next-generation sequencing (NGS) data [[83](#_ENREF_83)]. Coupled with community detection methods, this aims to identify subsets of organisms that work together typically for a specific task [[20](#_ENREF_20), [101](#_ENREF_101)]. Using these sub-communities identified within a larger interactome can provide a more resolved picture of ebbs and flows than phylum. For example, during tomato pomace deconstruction at 35C and at 55C, the dominant phylum in both unaerated conditions was Firmicutes but a Bray-Curtis analysis showed distinct differences between the communities (chapter 2). This would suggest that even though the phyla appear the same, differences in community function and subsequent OTUs are different. Applying network analysis to bio-solarization experiments, an organic amendment technique for soil, revealed distinct sub-communities that offered unique characteristics for each identified module (chapter 4).

While bio-solarization occurs in a field, field experiments a wrought with wide error bars due to lack of control of the system at hand. Using lab studies can offer a high degree of control but can have large disparities between what you see in the lab and what is seen in the field. It would be beneficial if you could use lab data to predict environmentally relevant sub-communities such that one may have an idea of what organisms or functional modules may arise under different conditions experienced in the field.

Predicting microbial communities can therefore utilize this interaction information and couple it with machine learning methods to provide estimates of microbial function and shape. Current prediction methods that utilize a DAG require generation of multiple graphs and select graphs that fit certain conditions [[46](#_ENREF_46)]. This can result in many trials of DAG generation prior to achieving a network that can fit within the requirements of this technique. A critical component to this prediction was to generate microbial interactions between the taxonomic level. This may also be accomplished through non-bayesian approaches by implementing an artificial neural network (ANN) with hidden layers. While the final equations to the DAG approximated a ANN, unique equations at each node were fitted by Eureqa.

This research aims to investigate the use of an interactome and an artificial neural network to predict community structure during the process of organic amendment of soil. This is critical to building the knowledge base around soil microbiomes and what they can contribute to carbon cycling and nitrogen cycling amongst other critical global cycles [[131](#_ENREF_131)]. This paper is the first known instance that attempts to predict microbiome structure using information based on the interactome.

## Methods

### Selection of Samples

Samples were selected based on their relevance to both soil and the organic amendment relevant in the studies used during the construction of the interactome.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Study Description | Number of Samples | Soil % (w/w) | Temperature (C) | Oxygen (%) | Duration (days) |
| Tomato Pomace Deconstruction |  | 0 | 35 – 55 | 0 – 21 | 7 – 21 |
| Tomato Pomace Solarization Controls |  | 93% – 100% | 25 – 55 | 0 – 21 | 8 – 21 |
| Mesophilic/Thermophilic Digestate Solarization Controls |  | 98.5% – 100% | 25 – 55 | 16 – 20 | 8 – 21 |

Table .: Sample metadata for construction of artificial neural network

### Artificial Neural Network Development and Error Clustering

The artificial neural network was developed with the core principles as described in Larsen and Gilbert, but not utilizing the DAG approach in favor of a traditional ANN with hidden nodes [[46](#_ENREF_46)]. Hidden layers are constructed using two unique layers; the first layer allows for pooling of information from sample metadata and the second layer allows for pooling of information for microbial modules. The first layer is only as big as is necessary to grasp potential interactions that explain the microbial community that can be predetermined by a CCA if available. The second layer is constrained by the number of microbial modules detected in the interactome. This ANN developed in this study contained four inputs that feed into a hidden layer with six perceptrons that eventually direct information into a second hidden layer with 11 perceptrons (supplementary figure).

The ANN was validated using leave-one-out where training occurred on n-1 samples and was tested against the remaining one sample. This was repeated for all n samples. Errors in prediction can follow systematic errors that can be analyzed to improve further prediction processes or establish limits of prediction. To understand these errors, an unsupervised classification algorithm implemented in R package MClust (version 5.2.2) [[132](#_ENREF_132)]. Bayesian information criterion (BIC) for selection of the most representative model for clustering.

### Community Prediction and Comparison to In-Field Solarization Samples

Prediction of the microbiome during components relevant for field conditions is critical to understand what is going on in between sampling. Using equations shown in Table 5.2, values were fed into the ANN developed including all samples to show predicted outputs over time. No appreciable degradation of organic matter was expected and outputs were recorded hourly.

|  |  |
| --- | --- |
| Description (variable) | Equation/Assumed Value |
| Time (T, days) | 21 |
| Temperature (temp, C) |  |
| Oxygen (oxygen, % atmosphere) |  |
| Soil (soil, % w/w) | 93% |

Table .: Parameters and Equations for modeling of in-field microbiome using ANN

In addition to raw output, data was normalized by equation X to account for proper estimation of proportions but incorrect raw values.

Using normalized relative abundance, predictions were made for potential field conditions based on three conditions. The shape of the ANN predicted communities was compared to field solarization samples using non-metric multidimensional scaling (NMDS) implemented in Vegan (version 2.4-2) (vegan citation). The input parameters for the following ANN predictions are shown in Table 5.3. Oxygen values were not recorded for the samples, but have been estimated from VFA production values by low activity (90% of normal atmosphere), medium activity (50% of normal atmosphere), and high activity (10% of normal atmosphere). Temperature was chosen as a mean of observed temperatures, with soil percentage and duration being the difference in percent amendment and when microcosms were extracted respectively.

|  |  |  |  |
| --- | --- | --- | --- |
| Sample | Soil | High Organic, Inoculum (-) | High Organic, Inoculum (+) |
| Soil (%) | 100 | 95 | 93 |
| Oxygen (% atmosphere) | 18.9 | 10.5 | 2.1 |
| Temperature (C) | 38 | 38 | 38 |
| Duration (days) | 8 | 8 | 8 |

Table .: Inputs to the ANN for NMDS comparison

## Results

### ANN biologic and training error

When training on all data, predicting the same data shows (Figure 5.1) the variation in the data due to biological inconsistencies. Most modules were expected to exist between 0 and 25% of relative abundance with only one module exceeding the 75% threshold; module 6.

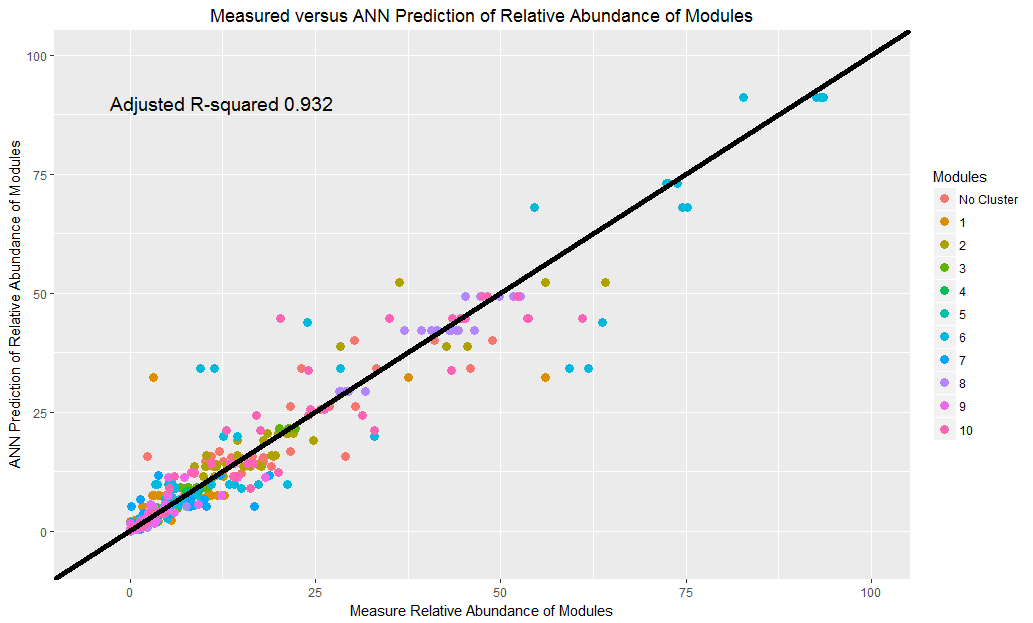


Figure .: Measured versus the prediction of ANN model

As seen in Figure 5.1, modules can have a lot of biologic variation, but by the nature of relative abundance all samples become interrelated. To that effect, investigating the interplay between errors by module can provide both insight into errors of predicting certain modules and how modules may be interrelated or interacting. MClust (version 5.2.2) was run on the leave-one-out cross-validation data where errors were recorded by module. MClust produced three distinct clusters (Table 5.3) with different distribution of errors among the modules (Figure 5.2). The characteristic of the three clusters were high, medium, and low error margins respectively. Most samples, 40 out of 57, were in the low – medium error clusters.

|  |  |  |
| --- | --- | --- |
| Distribution of Errors as Identified Using Ellipsoidal, Equal Orientation Clustering | | |
| Cluster 1 | Cluster 2 | Cluster 3 |
| 17 | 25 | 15 |

Table .: Number of samples in each error cluster identified by MClust

Figure 5.2 shows that modules 3 – 5 exhibited little to no error among the predictions. Modules 7 – 9 showed only minor errors across all error clusters. The largest errors were localized to modules 1, 6, and 10. Although modules 1 and 2 appear in the high error cluster, the high error is associated with two unique data points suggesting a potential outlier in the biologic data.

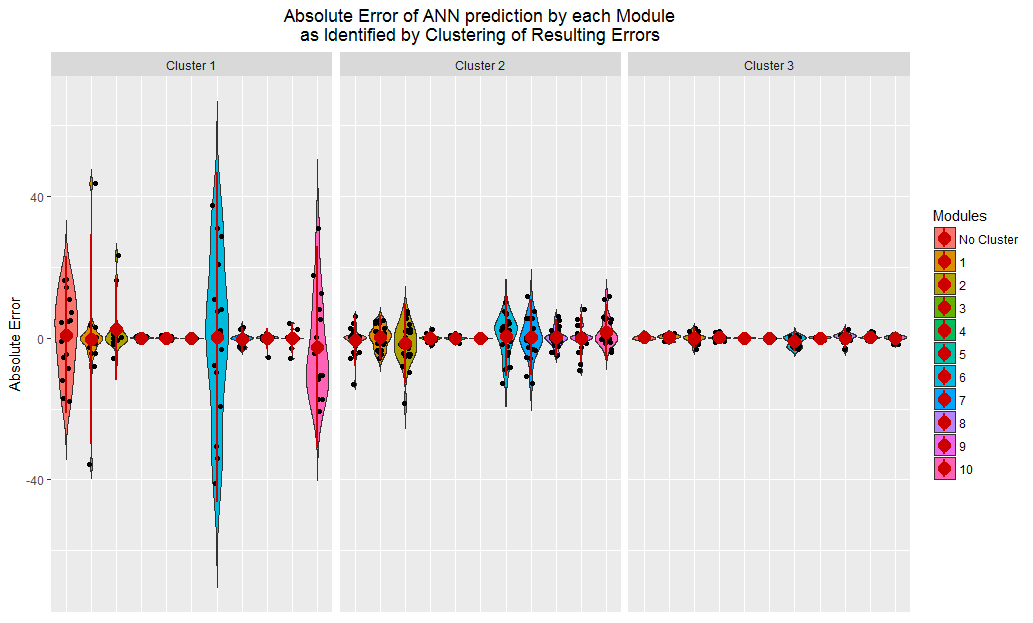


Figure .: Distribution of errors by cluster as shown by module

### Community prediction

Prediction of community dynamics remains a critical concept as it pertains to understanding microbiomes. Using the trained ANN, predictions were made regarding the distribution of modules throughout a 21-day amendment of organic matter that includes exponential decay of oxygen concentration in the soil, diurnal temperature cycling (black line; Figure 5.3) that also accounts for temperature accumulation during the solarization, percent of organic amendment in the soil, and duration. Raw ANN predictions (Figure 5.3A) appear to be capable of predicting distributions of relative abundance to near 100%. Deviations outside of this do occur suggesting more data may be necessary for better predictability. As the ANN is not constrained by its output, normalization of the output for relative abundance (Figure 5.3B) may show the appropriate proportions. As the raw ANN largely predicted values near 100% in total relative abundance, the effect of normalization was minor.

The initial two days show chaotic responses of module predictions associated with rapid changes in oxygen and temperature. Module 6 appears to enrich till approximately day 11 before tradeoffs between module 6 and 10 occur in coordination with temperature fluctuations. Module 10 appears initially enriched before decreasing over time till day 11 and then enriching till day 17 before a tradeoff between modules 1, 2, and 3 occurs. Rapid changes with temperature appear to be a second order response with oxygen and time a more first order in effect.

|  |  |
| --- | --- |
| A | B |

Figure .: Raw (A) and normalized (B) output from ANN predictions using equations in Table 5.2 The black line indicates current temperature associated with diurnal fluctuations as mimicked from Simmons et al. [[133](#_ENREF_133)].

Figure 5.4 shows the shape of both solarization samples from the field and ANN predicted samples based on measured temperatures, assumed oxygen concentrations, the organic amendment, and the day in which the sample was extracted. None of the field solarization samples were used in training or cross-validation of the data because of uncontrolled oxygen concentrations across all samples. This aims to understand prediction of community shape on which the interactome was established.

Each substrate appears to have its respective set of clustering within the NMDS with soil and the organically stable digestate appearing close together. The recalcitrant tomato pomace, both inoculated and un-inoculated, cluster together, and the wheat bran inoculated with compost clustered together with some minor overlap with tomato pomace samples. Predicted soil appears nestled between the stable digestate and soil samples but far away from all other amendments. The predicted organic amendment, inoculum (-) most resembled the samples from wheat bran with compost which the inoculum (+)

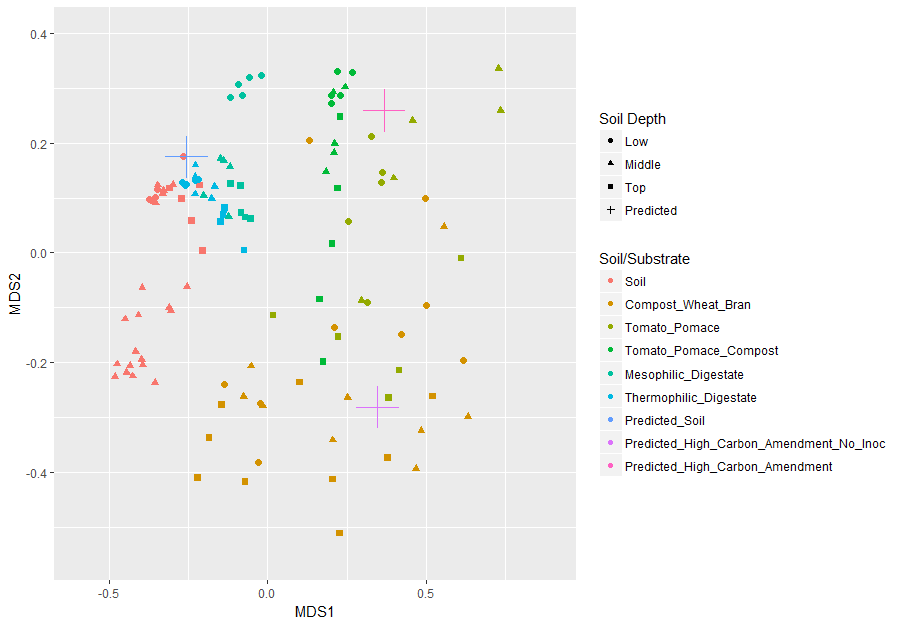


Figure .: NMDS of field solarization samples and ANN predicted communities

## Discussion

## Conclusion

# Concluding Remarks

Network too disconnected is useless

Network overly connected is useless

Network too sparse provides little information

Network too dense provides little information (or large false positives)

Stouffer’s p value combination [[134](#_ENREF_134)]

metaP (<https://cran.r-project.org/web/packages/metap/citation.html>)

P values 0<p<1 stouffers

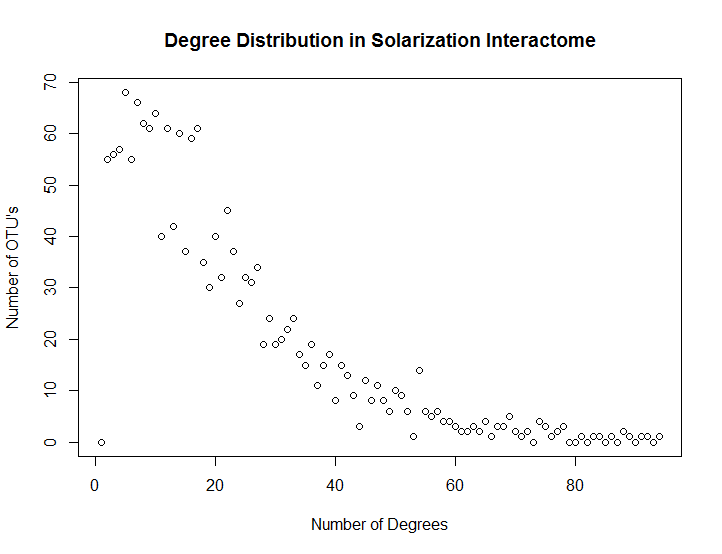
P values 0<p<=1 Fishers

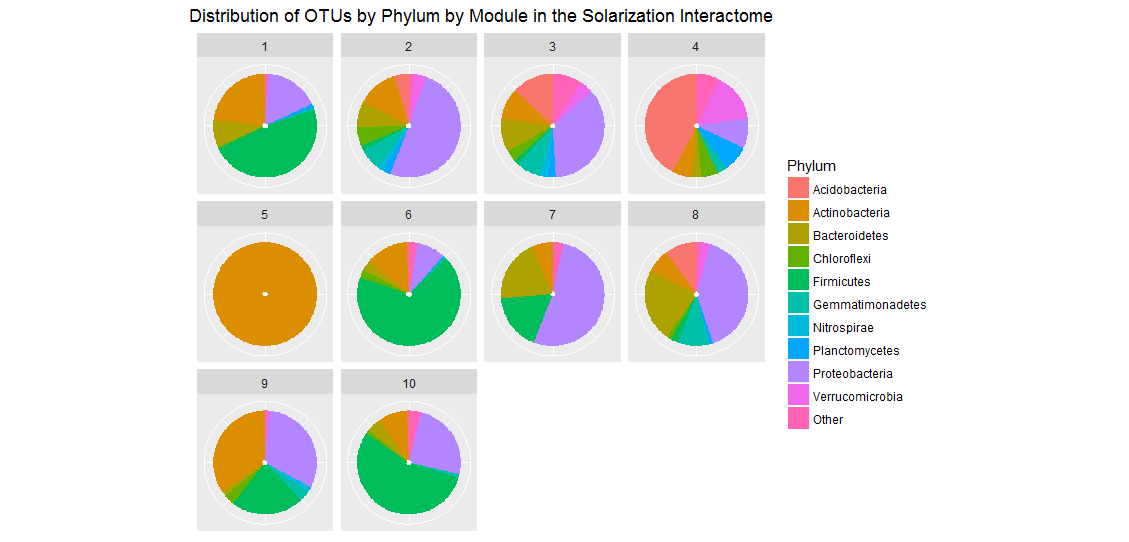
P values 0<=p<=1 sum of p [Becker, B J. Combining significance levels. In Cooper, H and Hedges, L V, editors A handbook of research synthesis, chapter 15, pages 215–230. Russell Sage, New York, 1994. Edgington, E S. An additive method for combining probability values from independent experiments. Journal of Psychology, 80:351-363, 1972.]

# APPENDIX 1

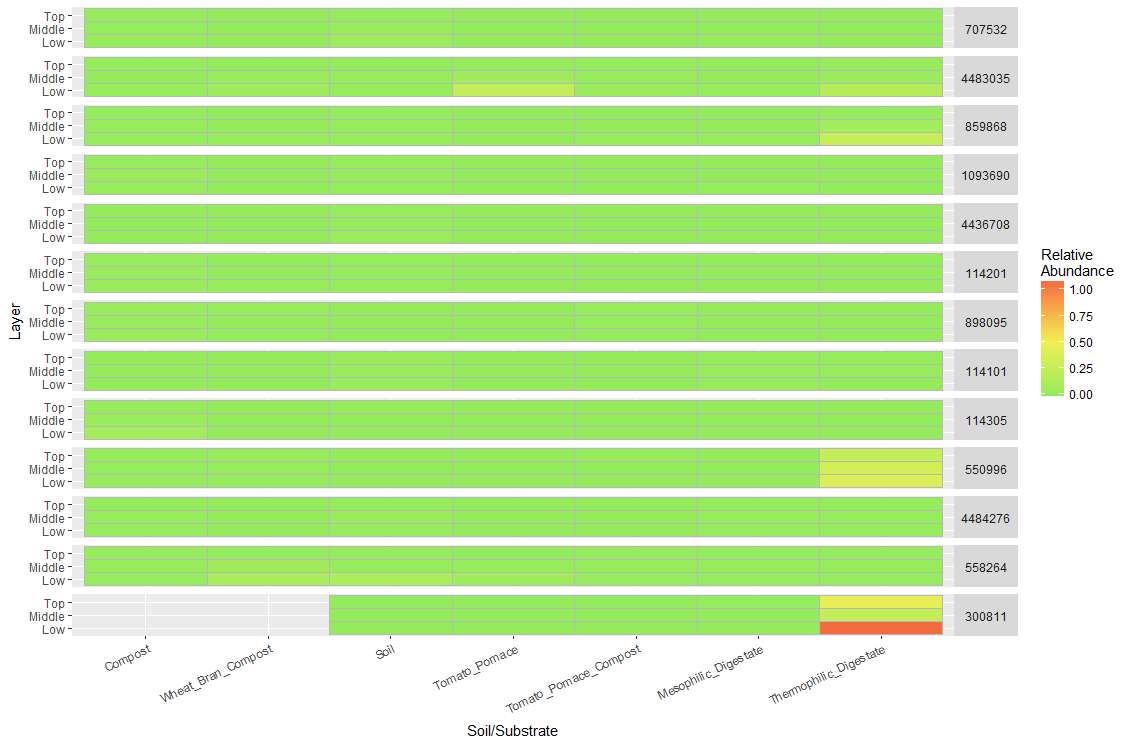
# APPENDIX 2

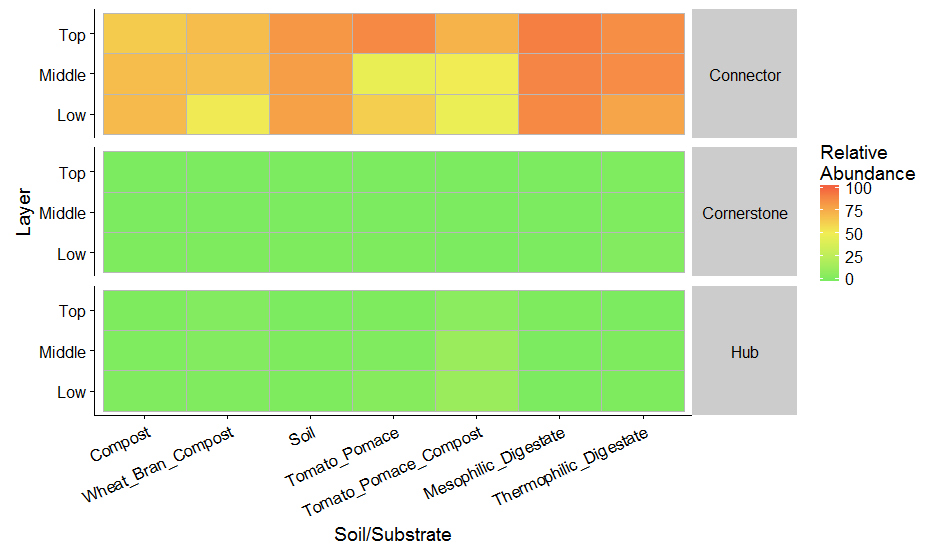
# APPENDIX 3

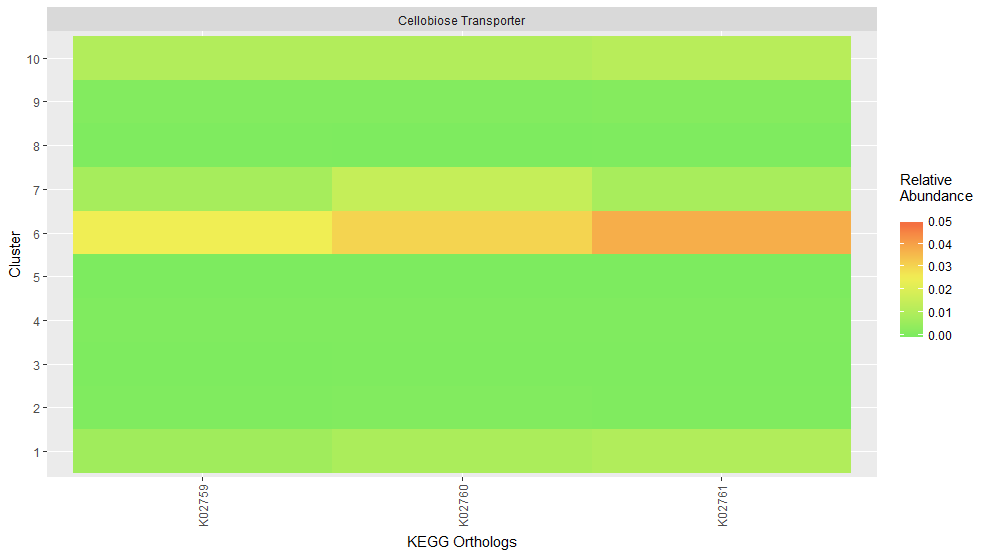


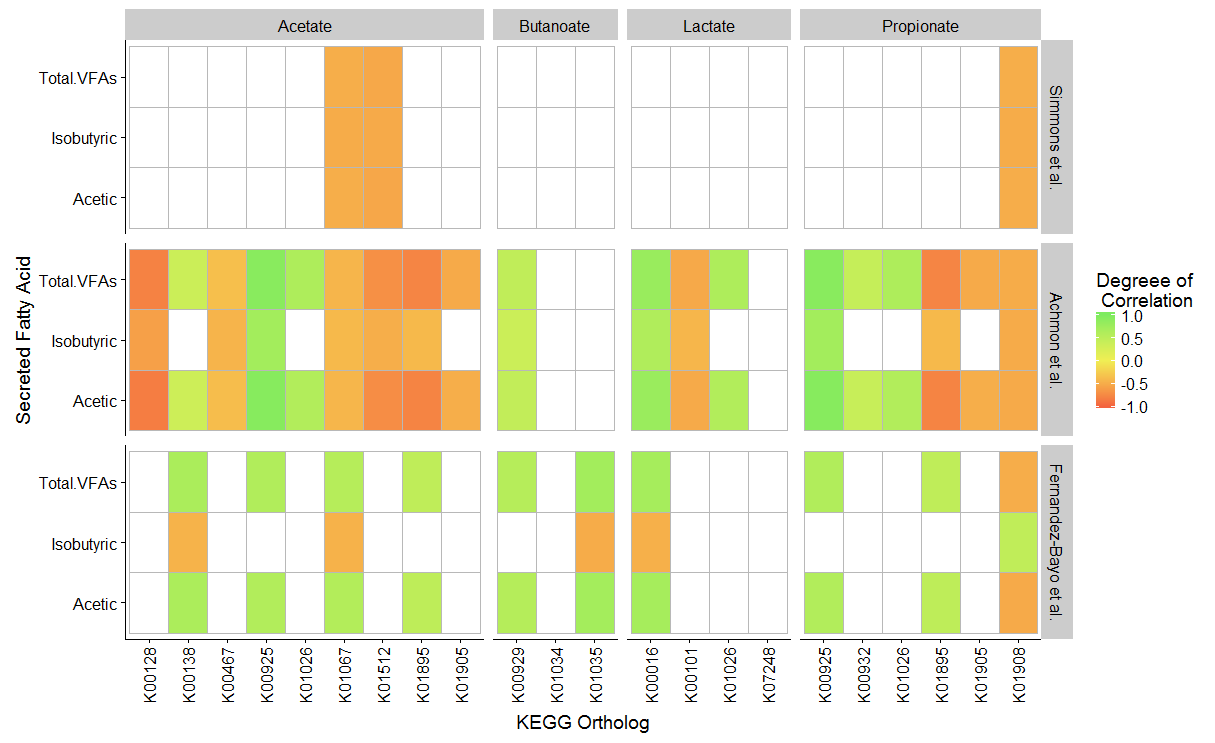


|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Cluster | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 |
| Number of OTUs | 155 | 250 | 316 | 97 | 2 | 273 | 91 | 261 | 62 | 176 |
| Number of Cornerstone | 4 | 0 | 3 | 0 | 0 | 0 | 0 | 4 | 2 | 0 |
| Number of Hubs | 1 | 3 | 7 | 1 | 0 | 10 | 1 | 3 | 0 | 4 |
| Number of Connectors | 130 | 158 | 205 | 80 | 0 | 114 | 61 | 199 | 60 | 129 |









# APPENDIX 4

# 

# References

1. Konopka, A., *What is microbial community ecology[quest].* ISME J, 2009. **3**(11): p. 1223-1230.

2. Clooney, A.G., et al., *Comparing Apples and Oranges?: Next Generation Sequencing and Its Impact on Microbiome Analysis.* PLoS ONE, 2016. **11**(2): p. e0148028.

3. Simmons, C., et al., *Metatranscriptomic analysis of lignocellulolytic microbial communities involved in high-solids decomposition of rice straw.* Biotechnology for Biofuels, 2014. **7**(1): p. 495.

4. Reddy, A.P., et al., *Discovery of Microorganisms and Enzymes Involved in High-Solids Decomposition of Rice Straw Using Metagenomic Analyses.* PLoS ONE, 2013. **8**(10): p. e77985.

5. DeAngelis, K.M., et al., *Characterization of Trapped Lignin-Degrading Microbes in Tropical Forest Soil.* PLoS ONE, 2011. **6**(4): p. e19306.

6. Großkopf, T. and O.S. Soyer, *Synthetic microbial communities.* Current Opinion in Microbiology, 2014. **18**(0): p. 72-77.

7. De Roy, K., et al., *Synthetic microbial ecosystems: an exciting tool to understand and apply microbial communities.* Environmental Microbiology, 2014: p. n/a-n/a.

8. Rui, J., et al., *The core populations and co-occurrence patterns of prokaryotic communities in household biogas digesters.* Biotechnology for Biofuels, 2015. **8**(1): p. 1-15.

9. Schimel, J. and S. Schaeffer, *Microbial control over carbon cycling in soil.* Frontiers in Microbiology, 2012. **3**(348).

10. Simmons, C.W., et al., *Characterization of bacterial communities in solarized soil amended with lignocellulosic organic matter.* Appl Soil Ecol., 2014. **73**.

11. Palmqvist, E. and B. Hahn-Hägerdal, *Fermentation of lignocellulosic hydrolysates. II: inhibitors and mechanisms of inhibition.* Bioresource Technology, 2000. **74**(1): p. 25-33.

12. Margot, J., et al., *Bacterial versus fungal laccase: potential for micropollutant degradation.* AMB Express, 2013. **3**: p. 63-63.

13. Lu, L., et al., *Environmental factors shaping the abundance and distribution of laccase-encoding bacterial community with potential phenolic oxidase capacity during composting.* Applied Microbiology and Biotechnology, 2015. **99**(21): p. 9191-9201.

14. Chen, Y., J.J. Cheng, and K.S. Creamer, *Inhibition of anaerobic digestion process: A review.* Bioresource technology, 2008. **99**(10): p. 4044-4064.

15. Reddy, A.P., et al., *Bioenergy feedstock-specific enrichment of microbial populations during high-solids thermophilic deconstruction.* Biotechnology and Bioengineering, 2011. **108**(9): p. 2088-2098.

16. Jiménez, D.J., D. Chaves-Moreno, and J.D. van Elsas, *Unveiling the metabolic potential of two soil-derived microbial consortia selected on wheat straw.* Scientific Reports, 2015. **5**: p. 13845.

17. Briones, A. and L. Raskin, *Diversity and dynamics of microbial communities in engineered environments and their implications for process stability.* Current Opinion in Biotechnology, 2003. **14**(3): p. 270-276.

18. Taha, M., et al., *Enhanced Biological Straw Saccharification Through Coculturing of Lignocellulose-Degrading Microorganisms.* Applied Biochemistry and Biotechnology, 2015. **175**(8): p. 3709-3728.

19. Liu, W., M. Pellegrini, and X. Wang, *Detecting Communities Based on Network Topology.* Scientific Reports, 2014. **4**: p. 5739.

20. Yang, Z., R. Algesheimer, and C.J. Tessone, *A Comparative Analysis of Community Detection Algorithms on Artificial Networks.* Scientific Reports, 2016. **6**: p. 30750.

21. Weiss, S., et al., *Correlation detection strategies in microbial data sets vary widely in sensitivity and precision.* ISME J, 2016. **10**(7): p. 1669-1681.

22. Berry, D. and S. Widder, *Deciphering microbial interactions and detecting keystone species with co-occurrence networks.* Frontiers in Microbiology, 2014. **5**(219).

23. Yang, B., Y. Wang, and P.-Y. Qian, *Sensitivity and correlation of hypervariable regions in 16S rRNA genes in phylogenetic analysis.* BMC Bioinformatics, 2016. **17**(1): p. 135.

24. Wang, Q., et al., *Naïve Bayesian Classifier for Rapid Assignment of rRNA Sequences into the New Bacterial Taxonomy.* Applied and Environmental Microbiology, 2007. **73**(16): p. 5261-5267.

25. Altschul, S.F., et al., *Basic local alignment search tool.* Journal of Molecular Biology, 1990. **215**(3): p. 403-410.

26. Ling, N., et al., *Insight into how organic amendments can shape the soil microbiome in long-term field experiments as revealed by network analysis.* Soil Biology and Biochemistry, 2016. **99**: p. 137-149.

27. Reddy, A.P., et al., *Thermophilic Enrichment of Microbial Communities in the Presence of the Ionic Liquid 1-ethyl-3-methylimidazolium acetate.* Journal of Applied Microbiology, 2012: p. n/a-n/a.

28. Siegwald, L., et al., *Assessment of Common and Emerging Bioinformatics Pipelines for Targeted Metagenomics.* PLOS ONE, 2017. **12**(1): p. e0169563.

29. Caporaso, J.G., et al., *QIIME allows analysis of high-throughput community sequencing data.* Nat Meth, 2010. **7**(5): p. 335-336.

30. Quast, C., et al., *The SILVA ribosomal RNA gene database project: improved data processing and web-based tools.* Nucleic Acids Research, 2012. **41**(D1): p. D590-D596.

31. Handelsman, J., et al., *Molecular biological access to the chemistry of unknown soil microbes: a new frontier for natural products.* Chemistry & Biology, 1998. **5**(10): p. R245-R249.

32. Wu, Y.-W., et al., *MaxBin: an automated binning method to recover individual genomes from metagenomes using an expectation-maximization algorithm.* Microbiome, 2014. **2**(1): p. 26.

33. Langille, M.G.I., et al., *Predictive functional profiling of microbial communities using 16S rRNA marker gene sequences.* Nat Biotech, 2013. **31**(9): p. 814-821.

34. Jimenez, D., F. Dini-Andreote, and J. van Elsas, *Metataxonomic profiling and prediction of functional behaviour of wheat straw degrading microbial consortia.* Biotechnology for Biofuels, 2014. **7**(1): p. 92.

35. Yu, C., et al., *Nitrogen amendment of green waste impacts microbial community, enzyme secretion and potential for lignocellulose decomposition.* Process Biochemistry.

36. Bowman, J.S. and H.W. Ducklow, *Microbial Communities Can Be Described by Metabolic Structure: A General Framework and Application to a Seasonally Variable, Depth-Stratified Microbial Community from the Coastal West Antarctic Peninsula.* PLOS ONE, 2015. **10**(8): p. e0135868.

37. Aßhauer, K.P., et al., *Tax4Fun: predicting functional profiles from metagenomic 16S rRNA data.* Bioinformatics, 2015. **31**(17): p. 2882-2884.

38. Yu, C., et al., *Nitrogen amendment of green waste impacts microbial community, enzyme secretion and potential for lignocellulose decomposition.* Process Biochemistry, 2017. **52**: p. 214-222.

39. Deng, Y., et al., *Molecular ecological network analyses.* BMC Bioinformatics, 2012. **13**(1): p. 113.

40. Layeghifard, M., D.M. Hwang, and D.S. Guttman, *Disentangling Interactions in the Microbiome: A Network Perspective.* Trends in Microbiology.

41. Aldecoa, R. and I. Marín, *Exploring the limits of community detection strategies in complex networks.* Scientific Reports, 2013. **3**: p. 2216.

42. Orman, G.K., V. Labatut, and H. Cherifi, *Comparative evaluation of community detection algorithms: a topological approach.* Journal of Statistical Mechanics: Theory and Experiment, 2012. **2012**(08): p. P08001.

43. Blondel, V.D., et al., *Fast unfolding of communities in large networks.* Journal of statistical mechanics: theory and experiment, 2008. **2008**(10): p. P10008.

44. Rosvall, M. and C.T. Bergstrom, *Maps of random walks on complex networks reveal community structure.* Proceedings of the National Academy of Sciences, 2008. **105**(4): p. 1118-1123.

45. Garcia, C., *BoCluSt: Bootstrap Clustering Stability Algorithm for Community Detection.* PLOS ONE, 2016. **11**(6): p. e0156576.

46. Larsen, P.E., D. Field, and J.A. Gilbert, *Predicting bacterial community assemblages using an artificial neural network approach.* Nat Meth, 2012. **9**(6): p. 621-625.

47. Langfelder, P. and S. Horvath, *Eigengene networks for studying the relationships between co-expression modules.* BMC Systems Biology, 2007. **1**: p. 54-54.

48. Demirbas, A., *Political, economic and environmental impacts of biofuels: A review.* Applied Energy, 2009. **86, Supplement 1**: p. S108-S117.

49. Matteson, G.C. and B.M. Jenkins, *Food and processing residues in California: Resource assessment and potential for power generation.* Bioresource Technology, 2007. **98**(16): p. 3098-3105.

50. Buchanan, B.B., et al., *Biochemistry and molecular biology of plants*. 2015: John Wiley & Sons.

51. Himmel, M.E., et al., *Biomass Recalcitrance: Engineering Plants and Enzymes for Biofuels Production.* Science, 2007. **315**(5813): p. 804-807.

52. Juturu, V. and J.C. Wu, *Insight into microbial hemicellulases other than xylanases: a review.* Journal of Chemical Technology & Biotechnology, 2013. **88**(3): p. 353-363.

53. Jørgensen, H., J.B. Kristensen, and C. Felby, *Enzymatic conversion of lignocellulose into fermentable sugars: challenges and opportunities.* Biofuels, Bioproducts and Biorefining, 2007. **1**(2): p. 119-134.

54. Klein-Marcuschamer, D., et al., *The challenge of enzyme cost in the production of lignocellulosic biofuels.* Biotechnol Bioeng, 2012. **109**.

55. Simmons, C.W., et al., *Effect of inoculum source on the enrichment of microbial communities on two lignocellulosic bioenergy crops under thermophilic and high-solids conditions.* J Appl Microbiol, 2014.

56. Yu, C., et al., *Preservation of microbial communities enriched on lignocellulose under thermophilic and high-solid conditions.* Biotechnology for Biofuels, 2015. **8**(1): p. 1-13.

57. Bokulich, N.A., et al., *Quality-filtering vastly improves diversity estimates from Illumina amplicon sequencing.* Nature methods, 2013. **10**(1): p. 57-59.

58. Parks, D.H., et al., *STAMP: statistical analysis of taxonomic and functional profiles.* Bioinformatics, 2014. **30**(21): p. 3123-3124.

59. Öztürk, M., *Conversion of acetate, propionate and butyrate to methane under thermophilic conditions in batch reactors.* Water Research, 1991. **25**(12): p. 1509-1513.

60. Thauer, R.K., K. Jungermann, and K. Decker, *Energy conservation in chemotrophic anaerobic bacteria.* Bacteriological Reviews, 1977. **41**(1): p. 100-180.

61. Siegert, I. and C. Banks, *The effect of volatile fatty acid additions on the anaerobic digestion of cellulose and glucose in batch reactors.* Process Biochemistry, 2005. **40**(11): p. 3412-3418.

62. Rahikainen, J., et al., *Inhibition of enzymatic hydrolysis by residual lignins from softwood—study of enzyme binding and inactivation on lignin-rich surface.* Biotechnology and Bioengineering, 2011. **108**(12): p. 2823-2834.

63. Lucey, K.S. and J.R. Leadbetter, *Catechol 2,3‐dioxygenase and other meta‐cleavage catabolic pathway genes in the ‘anaerobic’ termite gut spirochete Treponema primitia.* Molecular Ecology, 2014. **23**(6): p. 1531-1543.

64. Brown, M.E. and M.C.Y. Chang, *Exploring bacterial lignin degradation.* Current Opinion in Chemical Biology, 2014. **19**: p. 1-7.

65. van der Lelie, D., et al., *The Metagenome of an Anaerobic Microbial Community Decomposing Poplar Wood Chips.* PLoS ONE, 2012. **7**(5): p. e36740.

66. Merali, Z., et al., *Characterization of cell wall components of wheat straw following hydrothermal pretreatment and fractionation.* Bioresource Technology, 2013. **131**: p. 226-234.

67. Allison, B.J., et al., *The Effect of Ionic Liquid Pretreatment on the Bioconversion of Tomato Processing Waste to Fermentable Sugars and Biogas.* Applied Biochemistry and Biotechnology, 2016: p. 1-21.

68. Mhuantong, W., et al., *Comparative analysis of sugarcane bagasse metagenome reveals unique and conserved biomass-degrading enzymes among lignocellulolytic microbial communities.* Biotechnology for Biofuels, 2015. **8**(1): p. 1-17.

69. Wang, Y., et al., *A novel lignin degradation bacterial consortium for efficient pulping.* Bioresource Technology, 2013. **139**: p. 113-119.

70. Steinman, H.M., F. Fareed, and L. Weinstein, *Catalase-peroxidase of Caulobacter crescentus: function and role in stationary-phase survival.* Journal of Bacteriology, 1997. **179**(21): p. 6831-6.

71. Bandounas, L., et al., *Identification of a quinone dehydrogenase from a Bacillus sp. involved in the decolourization of the lignin-model dye, Azure B.* New Biotechnology, 2013. **30**(2): p. 196-204.

72. Podkaminer, K.K., et al., *Ethanol and anaerobic conditions reversibly inhibit commercial cellulase activity in thermophilic simultaneous saccharification and fermentation (tSSF).* Biotechnology for Biofuels, 2012. **5**(1): p. 1-9.

73. Yasir, M., et al., *Cloning and functional characterization of endo-β-1,4-glucanase gene from metagenomic library of vermicompost.* Journal of Microbiology, 2013. **51**(3): p. 329-335.

74. Wang, T., et al., *Directed evolution for engineering pH profile of endoglucanase III from Trichoderma reesei.* Biomolecular Engineering, 2005. **22**(1–3): p. 89-94.

75. Xing, D., et al., *Ethanoligenens harbinense gen. nov., sp. nov., isolated from molasses wastewater.* International Journal of Systematic and Evolutionary Microbiology, 2006. **56**(4): p. 755-760.

76. Ohmiya, K., et al., *Isolation and properties of beta-glucosidase from Ruminococcus albus.* Journal of Bacteriology, 1985. **161**(1): p. 432-434.

77. Del Valle, M., M. Cámara, and M.-E. Torija, *Chemical characterization of tomato pomace.* Journal of the Science of Food and Agriculture, 2006. **86**(8): p. 1232-1236.

78. Saini, J.K., et al., *Improvement of wheat straw hydrolysis by cellulolytic blends of two Penicillium spp.* Renewable Energy, 2016.

79. Neumüller, K.G., et al., *Synergistic action of enzyme preparations towards recalcitrant corn silage polysaccharides.* Biomass and Bioenergy, 2014. **60**: p. 88-97.

80. Xu, L., et al., *Biohydrogen production by Ethanoligenens harbinense B49: Nutrient optimization.* International Journal of Hydrogen Energy, 2008. **33**(23): p. 6962-6967.

81. Ortiz-Marquez, J.C.F., et al., *Genetic engineering of multispecies microbial cell factories as an alternative for bioenergy production.* Trends in Biotechnology, 2013. **31**(9): p. 521-529.

82. Zuroff, T.R. and W.R. Curtis, *Developing symbiotic consortia for lignocellulosic biofuel production.* Applied Microbiology and Biotechnology, 2012. **93**(4): p. 1423-1435.

83. Friedman, J. and E.J. Alm, *Inferring Correlation Networks from Genomic Survey Data.* PLOS Computational Biology, 2012. **8**(9): p. e1002687.

84. Kurtz, Z.D., et al., *Sparse and Compositionally Robust Inference of Microbial Ecological Networks.* PLOS Computational Biology, 2015. **11**(5): p. e1004226.

85. Xiong, W., et al., *Distinct roles for soil fungal and bacterial communities associated with the suppression of vanilla Fusarium wilt disease.* Soil Biology and Biochemistry, 2017. **107**: p. 198-207.

86. Aitchison, J., *The statistical analysis of compositional data.* 1986.

87. Chen, Z., *Is the weighted z‐test the best method for combining probabilities from independent tests?* Journal of Evolutionary Biology, 2011. **24**(4): p. 926-930.

88. Chen, Z., et al., *A new statistical approach to combining p-values using gamma distribution and its application to genome-wide association study.* BMC Bioinformatics, 2014. **15**(17): p. S3.

89. Corey, D.M., W.P. Dunlap, and M.J. Burke, *Averaging Correlations: Expected Values and Bias in Combined Pearson rs and Fisher's z Transformations.* The Journal of General Psychology, 1998. **125**(3): p. 245-261.

90. Guthman, J., *Going both ways: More chemicals, more organics, and the significance of land in post-methyl bromide fumigation decisions for California's strawberry industry.* Journal of Rural Studies, 2016. **47, Part A**: p. 76-84.

91. Ristaino, J.B. and W. Thomas, *Agriculture, Methyl Bromide, and the Ozone Hole: Can We Fill the Gaps?* Plant Disease, 1997. **81**(9): p. 964-977.

92. Yates, S.R., et al., *Emissions of 1,3-Dichloropropene and Chloropicrin after Soil Fumigation under Field Conditions.* Journal of Agricultural and Food Chemistry, 2015. **63**(22): p. 5354-5363.

93. Shennan, C., et al. *ANAEROBIC SOIL DISINFESTATION FOR SOIL BORNE DISEASE CONTROL IN STRAWBERRY AND VEGETABLE SYSTEMS: CURRENT KNOWLEDGE AND FUTURE DIRECTIONS*. 2014. International Society for Horticultural Science (ISHS), Leuven, Belgium.

94. Achmon, Y., et al., *Assessment of tomato and wine processing solid wastes as soil amendments for biosolarization.* Waste Management, 2016. **48**: p. 156-164.

95. Lazarovits, G., M. Tenuta, and K.L. Conn, *Organic amendments as a disease control strategy for soilborne diseases of high-value agricultural crops.* Australasian Plant Pathology, 2001. **30**(2): p. 111-117.

96. Abbasi, P.A., G. Lazarovits, and S. Jabaji-Hare, *Detection of High Concentrations of Organic Acids in Fish Emulsion and Their Role in Pathogen or Disease Suppression.* Phytopathology, 2009. **99**(3): p. 274-281.

97. Scherer, P. and L. Neumann, *“Methano-compost”, a booster and restoring agent for thermophilic anaerobic digestion of energy crops.* Biomass and Bioenergy, 2013. **56**: p. 471-478.

98. Riviere, D., et al., *Towards the definition of a core of microorganisms involved in anaerobic digestion of sludge.* ISME J, 2009. **3**(6): p. 700-714.

99. Poudel, R., et al., *Microbiome Networks: A Systems Framework for Identifying Candidate Microbial Assemblages for Disease Management.* Phytopathology, 2016: p. PHYTO-02-16-0058-FI.

100. Lancichinetti, A., S. Fortunato, and F. Radicchi, *Benchmark graphs for testing community detection algorithms.* Physical Review E, 2008. **78**(4): p. 046110.

101. Coutinho, F.H., et al., *Niche distribution and influence of environmental parameters in marine microbial communities: a systematic review.* PeerJ, 2015. **3**: p. e1008.

102. Franzén, O., et al., *Improved OTU-picking using long-read 16S rRNA gene amplicon sequencing and generic hierarchical clustering.* Microbiome, 2015. **3**: p. 43.

103. Olesen, J.M., et al., *The modularity of pollination networks.* Proceedings of the National Academy of Sciences, 2007. **104**(50): p. 19891-19896.

104. Guimera, R. and L.A. Nunes Amaral, *Functional cartography of complex metabolic networks.* Nature, 2005. **433**(7028): p. 895-900.

105. Humphries, M.D. and K. Gurney, *Network ‘Small-World-Ness’: A Quantitative Method for Determining Canonical Network Equivalence.* PLOS ONE, 2008. **3**(4): p. e0002051.

106. Telesford, Q.K., et al., *The Ubiquity of Small-World Networks.* Brain Connectivity, 2011. **1**(5): p. 367-375.

107. DeLong, E.F., et al., *Community Genomics Among Stratified Microbial Assemblages in the Ocean's Interior.* Science, 2006. **311**(5760): p. 496-503.

108. Faoro, H., et al., *Influence of Soil Characteristics on the Diversity of Bacteria in the Southern Brazilian Atlantic Forest.* Applied and Environmental Microbiology, 2010. **76**(14): p. 4744-4749.

109. Koyama, A., et al., *Soil bacterial community composition altered by increased nutrient availability in Arctic tundra soils.* Frontiers in Microbiology, 2014. **5**(516).

110. Coolen, M.J.L., et al., *Bioavailability of soil organic matter and microbial community dynamics upon permafrost thaw.* Environmental Microbiology, 2011. **13**(8): p. 2299-2314.

111. Mowlick, S., et al., *Changes and recovery of soil bacterial communities influenced by biological soil disinfestation as compared with chloropicrin-treatment.* AMB Express, 2013. **3**(1): p. 46.

112. Ju, F. and T. Zhang, *16S rRNA gene high-throughput sequencing data mining of microbial diversity and interactions.* Applied Microbiology and Biotechnology, 2015. **99**(10): p. 4119-4129.

113. McMurdie, P.J. and S. Holmes, *Waste Not, Want Not: Why Rarefying Microbiome Data Is Inadmissible.* PLOS Computational Biology, 2014. **10**(4): p. e1003531.

114. Brown, S.P., et al., *Scraping the bottom of the barrel: are rare high throughput sequences artifacts?* Fungal Ecology, 2015. **13**: p. 221-225.

115. Wang, W., et al., *Characterization of a microbial consortium capable of degrading lignocellulose.* Bioresource Technology, 2011. **102**(19): p. 9321-9324.

116. Koeck, D.E., et al., *First draft genome sequence of the amylolytic Bacillus thermoamylovorans wild-type strain 1A1 isolated from a thermophilic biogas plant.* Journal of Biotechnology, 2014. **192, Part A**: p. 154-155.

117. Wang, J.-Y., et al., *Intensive bioconversion of sewage sludge and food waste by Bacillus thermoamylovorans.* World Journal of Microbiology and Biotechnology, 2003. **19**(4): p. 427-432.

118. Combet-Blanc, Y., K.K. Kalamba, and P.Y. Kergoat, *Effect of pH on Bacillus thermoamylovorans Growth and Glucose Fermentation.* Applied and Environmental Microbiology, 1995. **61**(2): p. 656-9.

119. Maus, I., et al., *Unraveling the microbiome of a thermophilic biogas plant by metagenome and metatranscriptome analysis complemented by characterization of bacterial and archaeal isolates.* Biotechnology for Biofuels, 2016. **9**(1): p. 171.

120. Visser, M., et al., *Genome analysis of Desulfotomaculum kuznetsovii strain 17T reveals a physiological similarity with Pelotomaculum thermopropionicum strain SIT.* Standards in Genomic Sciences, 2013. **8**(1): p. 69-87.

121. Hausmann, B., et al., *Consortia of low-abundance bacteria drive sulfate reduction-dependent degradation of fermentation products in peat soil microcosms.* ISME J, 2016.

122. Carvalheiro, F., J.C. Roseiro, and M.T.A. Collaço, *Biological conversion of tomato pomace by pure and mixed fungal cultures.* Process Biochemistry, 1994. **29**(7): p. 601-605.

123. Martinez, D., et al., *Genome, transcriptome, and secretome analysis of wood decay fungus Postia placenta supports unique mechanisms of lignocellulose conversion.* Proceedings of the National Academy of Sciences, 2009. **106**(6): p. 1954-1959.

124. Kajikawa, H. and S. Masaki, *Cellobiose Transport by Mixed Ruminal Bacteria from a Cow.* Applied and Environmental Microbiology, 1999. **65**(6): p. 2565-2569.

125. Bengtson, P., et al., *Possible role of reactive chlorine in microbial antagonism and organic matter chlorination in terrestrial environments.* Environmental Microbiology, 2009. **11**(6): p. 1330-1339.

126. Leonowicz, A., et al., *Fungal laccase: properties and activity on lignin.* Journal of Basic Microbiology, 2001. **41**(3‐4): p. 185-227.

127. Oude Elferink, S.J.W.H., et al., *Anaerobic Conversion of Lactic Acid to Acetic Acid and 1,2-Propanediol by Lactobacillus buchneri.* Applied and Environmental Microbiology, 2001. **67**(1): p. 125-132.

128. Torino, M.I., et al., *Heterofermentative pattern and exopolysaccharide production by Lactobacillus helveticus ATCC 15807 in response to environmental pH.* Journal of Applied Microbiology, 2001. **91**(5): p. 846-852.

129. Chidthaisong, A. and R. Conrad, *Specificity of chloroform, 2-bromoethanesulfonate and fluoroacetate to inhibit methanogenesis and other anaerobic processes in anoxic rice field soil.* Soil Biology and Biochemistry, 2000. **32**(7): p. 977-988.

130. Kuang, J., et al., *Predicting taxonomic and functional structure of microbial communities in acid mine drainage.* ISME J, 2016. **10**(6): p. 1527-1539.

131. Trivedi, P., I.C. Anderson, and B.K. Singh, *Microbial modulators of soil carbon storage: integrating genomic and metabolic knowledge for global prediction.* Trends in Microbiology, 2013. **21**(12): p. 641-651.

132. Chris Fraley, A.E.R., *Model-Based Cluster, Discriminant Analysis and Density Estimation* Journal of American Statistical Association, 2002(97): p. 611-631.

133. Simmons, C.W., et al., *Managing compost stability and amendment to soil to enhance soil heating during soil solarization.* Waste Management, 2013. **33**(5): p. 1090-1096.

134. Whitlock, M.C., *Combining probability from independent tests: the weighted Z-method is superior to Fisher's approach.* Journal of Evolutionary Biology, 2005. **18**(5): p. 1368-1373.