Identification and Influence of Species-Informative 16S Ribosomal RNA Sequences and Evaluation of Ocean Biofilms

Nikhil Bose
University of Central Florida

Part of the Genetics and Genomics Commons

Find similar works at: https://stars.library.ucf.edu/etd2020
University of Central Florida Libraries http://library.ucf.edu

This Doctoral Dissertation (Open Access) is brought to you for free and open access by STARS. It has been accepted for inclusion in Electronic Theses and Dissertations, 2020- by an authorized administrator of STARS. For more information, please contact STARS@ucf.edu.

STARS Citation
IDENTIFICATION AND INFLUENCE OF SPECIES-INFORMATIVE 16S RIBOSOMAL RNA SEQUENCES AND EVALUATION OF OCEAN BIOFILMS

by

NIKHIL BOSE
B.S., BMS College of Engineering, 2014
M.S., University of California, Davis, 2016

A dissertation submitted in partial fulfillment of the requirements for the degree of Doctor of Philosophy in the Burnett School of Biomedical Sciences in the College of Medicine at the University of Central Florida Orlando, Florida

Spring 2023

Major Professor: Sean D. Moore
16S ribosomal RNA (rRNA) gene sequences are commonly analyzed for taxonomic and phylogenetic purposes because they contain variable regions that help distinguish genera. However, intra-genus classification is difficult due to high sequence similarity among closely related species. The biological impact of nucleotide variants in 16S variable regions are often unknown and hence their sequence differences are weighted evenly during classification, which provides poor species identity confidence. In this dissertation, I determined that analysis of intra-genus 16S allelic variants can provide species information and that nucleotide changes in 16S rRNA variable regions can impact ribosome quality. In one study, I analyzed ribosomal gene sequences, including 16S variable regions, to identify microbes that can spoil different retail draft beers. Based on relative sequence abundance changes of variable region sequences, I determined that certain bacteria preferred growth on draft lines rather than beers. Sequences of certain species were consistently detected at ratios indicative of their 16S gene copies, suggesting they came from specific strains. In a second study, I computationally interrogated 16S variable sequences in closely related genera Escherichia and Shigella and discovered that certain species could be differentiated. I demonstrated that Escherichia coli ribosomes were compromised when they carried 16S rRNA with these species-informative nucleotides, suggesting that variable region nucleotides may be constrained to respective species. In a third study, metagenomic sequencing was used to identify organisms that resided on cables submerged off the coast of Florida. Relative abundances of DNA for putative polymer-degrading organisms reduced over time and DNA for putative polymer-degrading enzymes were present at low relative abundance.
Altogether, this dissertation shows the capabilities of DNA-based microbial identification and suggests that acknowledgment 16S alleles can improve intra-genus bacterial classification.
ACKNOWLEDGMENTS

I thank Dr. Sean Moore for providing the opportunity to conduct this research and for mentoring me throughout my doctoral dissertation. I also thank Dr. Shibu Yooseph, Dr. Herve Roy, and Dr. Shaojie Zhang for their advice in my dissertation projects. Additional thanks go to Daniel Auvil, Anna Ward, and Laurie Agosto for their technical assistance in the beer project, and to Dr. Taj Azarian and Michael Johnstone for their editorial comments regarding publication of the tagged 16S project. The beer project was partially supported by NIH grant 1R01GM118896 and a UCF “What's Next” Quality Enhancement Plan grant, and the tagged 16S project was supported by NIH grant 1R01GM118896.
# TABLE OF CONTENTS

<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>ABSTRACT</td>
<td>iii</td>
</tr>
<tr>
<td>ACKNOWLEDGMENTS</td>
<td>v</td>
</tr>
<tr>
<td>TABLE OF CONTENTS</td>
<td>vi</td>
</tr>
<tr>
<td>LIST OF FIGURES</td>
<td>ix</td>
</tr>
<tr>
<td>LIST OF ACRONYMS (or) ABBREVIATIONS</td>
<td>x</td>
</tr>
<tr>
<td>CHAPTER ONE: INTRODUCTION</td>
<td>1</td>
</tr>
<tr>
<td>Spoilage Microbes in Retail Draft Beers and the Biofilms they form</td>
<td>2</td>
</tr>
<tr>
<td>16S Ribosomal RNA Genes may be Tailoring their Variability to Evolution</td>
<td>3</td>
</tr>
<tr>
<td>Shotgun Metagenomics to Identify Putative Polymer-degrading Organisms and Enzyme Genes</td>
<td>5</td>
</tr>
<tr>
<td>CHAPTER TWO: LITERATURE REVIEW</td>
<td>6</td>
</tr>
<tr>
<td>16S rRNA Variable Region Analysis</td>
<td>6</td>
</tr>
<tr>
<td>Computational Processing of Ribosomal Gene Sequencing Data</td>
<td>8</td>
</tr>
<tr>
<td>Apparent Disconnect Between 16S rRNA Nucleotide Conservation from Phylogenetic, Taxonomic, Structural, and Functional Evaluations</td>
<td>9</td>
</tr>
<tr>
<td>Metagenomics Evaluation of Biofilms Formed on Polymers Submerged in the Ocean</td>
<td>12</td>
</tr>
<tr>
<td>CHAPTER THREE: METHODOLOGY</td>
<td>13</td>
</tr>
<tr>
<td>Beer Cultures and Microbial DNA Analysis</td>
<td>13</td>
</tr>
<tr>
<td>Beer Sample Collection and Culturing</td>
<td>13</td>
</tr>
<tr>
<td>DNA Extraction from Starter Stocks and Beer Cultures</td>
<td>14</td>
</tr>
<tr>
<td>DNA Library Preparation and Sequencing</td>
<td>16</td>
</tr>
<tr>
<td>Sequence Processing and Classification of Microbial Organisms in Beer</td>
<td>16</td>
</tr>
<tr>
<td>Taxonomy Bush and Graphics</td>
<td>17</td>
</tr>
<tr>
<td>Variable Region rRNA Sequences Influence Ribosome Performance</td>
<td>18</td>
</tr>
<tr>
<td>Strains and Plasmids</td>
<td>18</td>
</tr>
<tr>
<td>Experimental Culturing, Harvesting, and Lysis</td>
<td>18</td>
</tr>
<tr>
<td>Cell Fractionation</td>
<td>19</td>
</tr>
<tr>
<td>RNA Extraction</td>
<td>20</td>
</tr>
<tr>
<td>RT-qPCR</td>
<td>21</td>
</tr>
<tr>
<td>Retrieval of 16S Gene V1 and V3-V4 Sequences and Relative Entropy Analysis</td>
<td>21</td>
</tr>
</tbody>
</table>
V3 Sequence Determination for Gammaproteobacteria and Clostridioides difficile ........ 23
Structure Analysis ............................................................................................................. 23
Polymer Cable Sample Processing ........................................................................................................ 24
Cables Jacket Materials and their Retrieval from the Ocean ................................................................. 24
Seawater Retrieval ......................................................................................................................... 24
DNA extraction from Cable Samplings and Seawater ........................................................................... 24
DNA Quantitation, Normalization, and Sequencing ............................................................................... 25
Bioinformatics: Sequence Processing ................................................................................................. 26
Bioinformatics: Genus Classification and Relative Abundance Estimation of Polymer-degrading Genera ......................................................................................................................... 26
Bioinformatics: Classification of Polymer-degrading Enzymes ............................................................. 27
CHAPTER FOUR: FINDINGS .............................................................................................................. 28
Microbial Analysis of Retail Draft Beers ............................................................................................. 28
Establishing a Test Platform for Beer Microbiota. ................................................................................ 28
Identifying Bacteria and Fungi ........................................................................................................... 30
Microbial Diversity in the Starter Samples ........................................................................................... 30
Dominant Culturable Microbes .......................................................................................................... 32
Preferences for Biofilm or Planktonic Growth .................................................................................... 34
Ribosome Abundance Changes from 16S rRNA Mutations .................................................................. 39
Development of a Tagged 16S rRNA Tracking System ....................................................................... 39
Distribution of Toxic 16S rRNA Mutants .............................................................................................. 43
Disparate V3 Region Sequences affect 16S rRNA Abundances in Ribosomes .................................... 44
Positional Relative Entropy Reveals Strain- and Species-specific Residue Variations ....................... 44
Identification of Strain- and Species-informative rRNA Variants ......................................................... 49
Escherichia and Shigella V3-V4 Allele Variants Influence Ribosome Performance ............................. 51
Microbial Analysis of Polymer Cables Submerged in the Ocean ......................................................... 54
Establishing a Method to Extract DNA from Cable Biofilms and Seawater ........................................ 54
DNA Read Classifications and their Relative Abundances for Biofilm on Unsubmerged Cables ........................................................................................................................................... 55
Relative Abundances of Classified DNA for Submerged Cables and Surrounding Water .................. 57
A Focus on Genera for Putative Polymer-degrading Species on Cables ............................................. 61
A Focus on Gene Sequences that Corresponded to Polymer-degrading Enzymes .............................. 63
CHAPTER FIVE: CONCLUSION................................................................. 69
APPENDIX A: COPYRIGHT INFORMATION FOR BEER LINE BIOFILM STUDY ........ 74
APPENDIX B: SPECIFICITY EVALUATION OF qPCR PRIMERS FOR TAGGED V1 16S 76
APPENDIX C: E. COLI CULTURE GROWTH RATES FOR DIFFERENT PLASMID EXPRESSIONS (WITH CONTROLS) ........................................................................................................... 78
APPENDIX D: PERFORMANCE ASSESSMENT OF PARENTAL 16S FOR VARIOUS EXPRESSION CONDITIONS ......................................................................................... 80
APPENDIX E: E. COLI AND C. DIFF V3 REGION ALIGNMENT............................. 82
APPENDIX F: BETA DIVERSITY COMPARING WATER AND CABLE SAMPLES .......... 84
LIST OF REFERENCES .............................................................................. 86
LIST OF FIGURES

Figure 1: Beer sampling and biofilm development ................................................................. 29
Figure 2: Bacteria and fungi present in the starter samples .................................................... 32
Figure 3: Diversity of bacterial genera and their preferential growth ..................................... 33
Figure 4: Evaluating bacterial biofilm preferences ................................................................. 36
Figure 5: Structure and sequence analysis of E. coli 16S rRNA V1 region ............................ 40
Figure 6: Establishing the abundances of 16S rRNA variants .............................................. 42
Figure 7: Evaluating decoding center mutants ........................................................................ 43
Figure 8: Identification and performance assessment of disparate V3 region variants .......... 46
Figure 9: Illustration of V3-V4 positional relative entropy ..................................................... 48
Figure 10: Informative V3-V4 sequence polymorphisms among Escherichia and Shigella ...... 51
Figure 11: Structure of V3-V4 informative residues in E. coli and abundance scores for
Escherichia and Shigella species variants ............................................................................. 53
Figure 12: Process for DNA extraction from ocean-submerged cables .................................. 55
Figure 13: Relative Abundances of DNA for Genera detected on Unsubmerged Cables ....... 57
Figure 14: Total Relative Abundances of Classified DNA for Submerged Cables and Water
Samples .................................................................................................................................. 59
Figure 15: Average Relative Abundance (RA) of Genera for Putative Polymer-degrading
Organisms .............................................................................................................................. 62
Figure 16: Total Relative Abundances of Putative Polymer-degrading Enzymes .................. 65
Figure 17: Average Relative Abundance (RA) of DNA for Polymer-degrading Enzymes ....... 67
LIST OF ACRONYMS (or) ABBREVIATIONS

CG: Clustered visible Growth
DNA: Deoxyribonucleic Acid
HDPE: High Density Polyethylene
ITS: Internal Transcribed Spacers
LG: Low visible Growth
OTU: Operational Taxonomic Unit
PCR: Polymerase Chain Reaction
PP: Polypropylene
PVC: Polyvinyl Chloride
PVDF: Polyvinylidene Fluoride
RA: Relative Abundance
RNA: Ribonucleic Acid
rRNA: Ribosomal Ribonucleic Acid
SNP: Single Nucleotide Polymorphism
TPE: Thermoplastic Elastomer
V1 region: Variable Region 1 of the 16S rRNA gene
V3-V4 region: Variable Regions 3 and 4 of the 16S rRNA gene
VG: Visible Growth
zOTU: zero-radius Operational Taxonomic Unit
CHAPTER ONE: INTRODUCTION

Historically, microbial species were identified based on observations under a microscope and biochemical tests, which relied on the ability to culture an isolated organism (1). A quicker and robust method in modern times involves matching DNA sequences derived from a microbial community to genomes of species or strains (2). DNA sequences can be analyzed for classification in one of two ways, a) by sequencing specific PCR-amplified genes (called amplicon sequencing), which contain taxonomically informative variations among organisms, or b) by sequencing all DNA fragments (called metagenomic sequencing) and classifying those that are indicative of specific organisms (3). While metagenomic sequencing is useful for classifying multiple genes, modern sequencing methods to do so are expensive and sequence analysis strategies are computationally cumbersome. Alternatively, amplicon sequencing is economically and computationally less cumbersome, and genes analyzed are often provide genus-level classification (4).

Ribosomal RNA (rRNA) genes are commonly targeted for amplicon sequencing because they are present in all organisms and they contain sequence differences that help differentiate organisms (5). The 16S rRNA gene is most commonly analyzed for bacterial and archaeal classification (6), and the 18S rRNA gene or internal transcribed spacers (ITS) of ribosomal genes are generally analyzed for eukaryotic classification (7). Taxonomically informative 16S rRNA gene variations are often present in nine regions called variable regions (V1-V9). These variable regions are commonly PCR-amplified using primers that bind to flanking regions of sequence conservation across multiple microbes (8). The combined analysis of 16S variable regions 3 and 4 (V3-V4 region) is compatible with affordable second-generation sequencing.
platforms and results often provide genus-level, or sometimes, species-level identification (9). However, species classification often comes with lower confidence due to high similarity among closely related species.

The utility of strain- or species-level 16S rRNA allelic variants to improve taxonomic classification is often unexplored because the biological relevance of the few nucleotide differences is often unknown. In this dissertation, allelic variants of the 16S rRNA V3-V4 region sequences were explored to obtain species-level information and the biological impact of species-informative variants were evaluated. Also, the utility of metagenomic sequencing to classify organisms and polymer-degrading enzyme genes was determined for polymer cables submerged in the ocean. These studies re-evaluate the notion that 16S gene variable region sequences are uninformative for intra-genus classification and that single nucleotide variations within them have no consequence to strains that bear them (10, 11). The strategies for microbial identification using amplicon and metagenomic sequencing are described in a cogent manner for appropriate genus- or species-level classification. Furthermore, this dissertation identifies microbes and polymer-degrading enzymes on highly durable polymers retrieved from ocean environments, which is often not investigated in biofouling studies.

**Spoilage Microbes in Retail Draft Beers and the Biofilms they form**

Beer production involves controlled fermentation of sugar extracts by yeasts in the presence of water and hops, such that the introduction of unwanted microbes can ruin beer quality. A previous study determined sources of beer spoilage microbes in various parts of the brewing process (12), however draft-dispensed beer in retail environments is typically not monitored. The growth of beer spoilage microbes in beer draft lines due to inappropriate
cleaning can deter beer quality, which lowers customer satisfaction and economic stability of the retail establishment. Amplicon sequencing serves as an affordable method to monitor beer spoilage microbes and their abundances in draft lines and beers.

In this study, amplicon sequencing of the 16S rRNA V3-V4 and the ITS2 region was employed to identify spoilage bacteria and fungi in retail draft beers. Four different beers were sampled in two separate years from a retail environment and cultured in beer-medium containing draft line PVC under laboratory-controlled to evaluate biofilm vs planktonic growth. Results showed that up to 119 bacterial and 18 fungal species inhabited beers and retail draft lines. Several bacteria preferably grew in biofilms on draft lines, suggesting that they may persist and contaminate fresh beers if draft lines are not cleaned appropriately. Several V3-V4 sequences that classified to the same species were at ratios indicative of 16S gene allele variants from the same strain, which bolstered confidence in their identification. This study therefore provides a method to monitor spoilage microbes in beer draft lines and draft beers, as well as suggest improvements for rRNA sequence classification by acknowledging intra-species allelic variants.

**16S Ribosomal RNA Genes may be Tailoring their Variability to Evolution**

Although 16S rRNA variable region sequences have diverged unevenly across bacteria (6), computational tools used for classification have weighted nucleotide differences among sequences evenly (13, 14). Furthermore, 16S genes often occur as multiple copies per genome, each with potential nucleotide differences that are typically considered uninformative in differentiating species (15). Because the biological importance of 16S variable region nucleotides and nucleotide variants are often unknown, relatedness within a genus is commonly
assigned by overall sequence similarity and the potential for individual residue identities across multiple 16S gene alleles to indicate considerably larger evolutionary divergence is overlooked.

In this study, an in vivo system was developed where a 16S rRNA variant was expressed from a plasmid in *Escherichia coli* that also expressed chromosome-born 16S rRNA and their relative abundance determined in stages of small subunit assembly, ribosome assembly, and translation. Notably, an SNP in the V3 region led to 16S variant abundances in ribosomes as low as that observed for a truncated variable region 3 (V3) as seen in *Clostridioides difficile*, which suggests that certain variable region nucleotides in *E. coli* are constrained from mutating. 16S variable region sequences in the genera *Escherichia* and *Shigella* were then computationally evaluated to identify single nucleotide variants that could distinguish certain species. Results showed that pathogenic *E. albertii, S. boydii*, and *S. dysenteriae* species could be distinguished from *E. coli* based on the same nucleotide variations observed consistently across 16S gene copies in their respective strains. *E. coli* 16S rRNA harboring these species-informative variants showed compromised abundances in ribosomes, suggesting a biological impact from variable region mutation. The only variant that demonstrated uncompromised abundances in ribosomes was a four-nucleotide covariation seen in *E. albertii*, which suggests that multiple evolutionary steps were required to maintain ribosome quality. Overall, this study revealed that variable region nucleotide polymorphisms are not necessarily inconsequential and that computational approaches for 16S variable region sequence classification should not assume an even probability of residue divergence.
**Shotgun Metagenomics to Identify Putative Polymer-degrading Organisms and Enzyme Genes**

While the economic and computational costs for metagenomics may be more than amplicon sequencing, the ability to analyze multiple gene sequences all-at-once makes metagenomic sequencing more efficient than separately amplicon sequencing individual genes. Also, species or genus classification can be more reliable due to simultaneous classification of multiple DNA fragments (3, 16). Metagenomic sequencing was recently pursued in bioremediation studies to identify organisms capable of degrading plastics and their genes that express plastic-degradation enzymes (17–19). While some studies have identified microbes responsible for biodegradation of polyester (PE) and polypropylene (PP), highly durable polymers such as thermoplastic elastomer (TPE) and polyvinylidene fluoride (PVDF) have not been explored.

In this study, shotgun metagenomics was used to identify biofilm organisms and polymer-degradation enzymes that were present on cable outer jackets made of TPE, PVDF, HDPE, or PP. These cables were submerged in the ocean and the biofilms that formed were investigated after certain timepoints. With increased submersion time, cables exhibited decreased relative abundances (RA) of classified organisms, including putative polymer-degrading organisms. Genera for putative polymer-degrading organisms that were consistently detected across all cables included *Sphingomonas, Pseudomonas, Streptomyces, Vibrio,* and *Candida* RA for polymer-degrading enzymes varied across timepoints on all cables but were always less than 0.15%. Ultimately, the cable jackets tested in this study showed low dominance of DNA for putative polymer-degradation organisms and enzymes, and that TPE may be a preferred choice for long-term submersion due to consistently low RA for enzyme genes.
CHAPTER TWO: LITERATURE REVIEW

Early works of characterizing microbes were performed in 1860, when Louis Pasteur used controlled experiments to determine that yeasts were responsible for alcohol fermentation (20). Also, in the late 1800s, the work of Robert Koch (21) paved the way for methods to isolate bacteria from environments and the work of Hans Christian Gram (22) paved the way for differentiating bacteria using chemical tests. Since then, different selective growth media have been used to isolate various microbes, after which they are characterized based on several biochemical tests (1, 23, 24). However, isolating every individual species from an environment is extensively laborious and may not be efficient in retrieving all species. In 1987, Carl Woese first proposed the analysis of the ribosomal RNA (rRNA) genes for phylogeny because their sequences differed among bacteria and archaea (5, 25). He noted that genes for the 16S rRNA and 23S rRNA were present in all bacteria and archaea and that their sequences served as “molecular chronometers” that accumulated variations as the organism evolved (5). Accompanied by the advent of polymerase chain reaction (PCR) in the 1980s (26), gene sequences could be amplified and analyzed to characterize the DNA makeup of organisms in microbial communities, which avoided culture-based isolation altogether.

16S rRNA Variable Region Analysis

Based on the work of Carl Woes, the 16S gene could be analyzed for phylogeny and taxonomy of bacteria (5). Van de Peer et al. mapped the substitution rate of 16S sequences across known sequences and identified nine regions of sequence variability in a prokaryotic secondary structure (V1-V9) (27). Baker et al. then identified regions in the 16S gene with high sequence conservation among bacteria where oligonucleotide primers could bind for PCR amplification of
16S gene segments (8). A conserved region flanked each of the nine variable regions and they were mapped to an *E. coli* 16S gene as a reference, which is still used to date. The sequences of these nine variable regions could be analyzed to distinguish bacteria for phylogenetic and taxonomic purposes and the more variable regions analyzed in tandem, the greater the taxonomic resolution to distinguish closely-related species (6, 28).

Though the current notion is that no individual variable region is reliably capable of classifying all bacteria compared to the full length 16S sequence (6, 29), methods to sequence the full length 16S gene are expensive. As such, researchers have debated over which variable regions can classify the most bacteria, with the V3, V4, and V6 regions showing popularity for detecting species or genera (30–33). Furthermore, though segments flanking the variable regions are often considered "conserved", slight variations in sequences have been observed among bacteria, which has led to frequent re-evaluations of appropriate primer sequences that bind to 16S genes for most bacteria and archaea (34–37). An extensive study evaluating primers, establish a primer set that flanked the V3 and V4 regions of ~96% of bacteria and ~65% of archaea (38) in a database consisting of several thousands of non-redundant 16S sequences. Furthermore, the amplicons generated using this primer set could then be used in affordable, high accuracy, short-read second-generation sequencing platforms (9, 28, 39). The V3 and V4 regions of 16S genes have hence been extensively analyzed for bacterial classification in pathogenic and environment samples (9, 34, 39–44). Due to the sequencing length limit of affordable second-generation sequencing methods, eukaryotic internal transcribed spacer (ITS) regions have also been analyzed for taxonomic purposes (12, 28).

A previous amplicon sequencing study identified beer-spoilage microbes associated with different parts of a brewery by analyzing 16S V4 regions for bacteria and ITS regions for fungi
This brewery study identified families, and in some cases genera, of beer spoilage microbes, which if present in raw material (grain, yeast, hop, and recycled beer) were likely contributors of contamination in other parts of the brewery including casks of the produced beer. As such, sequencing 16S variable regions and ITS regions on second-generation sequencing platforms is suitable for identifying beer spoilage microbes. However, the microbial monitoring in retail draft beer is often not done and hence the communities that establish are unknown.

Computational Processing of Ribosomal Gene Sequencing Data

The raw data generated from second generation sequencing can be several gigabytes per sample. Therefore, computational tools, such as USEARCH (13), QIIME2 (14), and mothur (45), have been developed that incorporate efficient codes and pipelines that expediently provide information on microbial classification and community diversity. 16S gene sequences are classified by identifying near-closest matches to reference sequences in databases annotated with genus and species names of organisms. However, prior to classification, one of two possible sequence clustering strategies may be applied, operating under different assumptions. The first strategy involves clustering unique sequences that are at least 97% similar under a single representative sequence labelled as an operational taxonomic unit (OTU) (45, 46). In this case, the OTU sequence is considered to be species-representative, and all sequences clustered under it as variants of that species. However, the 97% sequence identity does not always distinguish all species or even genera (28, 47). For example, 16S gene sequences for strains under the genera Escherichia and Shigella are more than 97% similar, making their genera indistinguishable (47, 48). Alternatively, the second strategy assumes that each unique sequence may represent a separate organism and are hence classified (these unique sequences are called Amplicon
Sequence Variants [ASVs] or zero-radius Operational Taxonomic Units [zOTUs]) (49, 50).

However, zOTUs often yield the same classifications for multiple sequences because bacterial species often have multiple copies of the 16S gene in their genome, each of which may have 1-4 sequence differences (15, 51). Also, classification can be at a lower confidence for variable regions when sequences are similar to references of multiple species (6, 28). A partial solution for improved classification confidence may be to add more species variant references into databases and apply stringent match criteria, however low classification confidence may still occur when closely related species share one or several identical 16S alleles. One additional solution could be to determine if sequences with redundant classifications are indicative of genomic copies from the same species (52, 53). However, such an attempt has not previously been explored for 16S variable region sequence analysis.

Apparent Disconnect Between 16S rRNA Nucleotide Conservation from Phylogenetic, Taxonomic, Structural, and Functional Evaluations

16S rRNA sequence analysis for phylogenetic, taxonomic, or structural evaluation purposes is often disconnected from the conservation of 16S rRNA nucleotides among bacteria. Carl Woese’s report on bacterial evolution stated that different nucleotides in ribosomal RNA genes mutate at different rates (5), which suggested that distantly related microbes have residue differences at positions with infrequent changes and closely related microbes have residue differences at positions with frequent changes. However, nucleotide conservation in 16S rRNA genes is currently segregated into specific “regions” even though nucleotides in these regions have not similarly varied across all bacteria. Initial reports that demarcated conserved 16S rRNA regions were based on identification of oligonucleotide primer-binding sites in 16S genes for a set of bacteria under study, while regions that were incompatible with primers were deemed as
variable (8). However, evidence from several studies that followed demonstrated incompatibility of primers with 16S sequences of certain bacteria due to nucleotide mismatches (36–38).

Furthermore, Shannon entropy analysis of 16S rRNA sequences for bacteria showed that variable region nucleotides do not diverge evenly (6, 29). Therefore, not all “conserved region” nucleotides have been evolutionarily conserved across bacteria, and not all “variable region” nucleotides have freely varied. As such, evolutionary divergence among strains in a species may not necessarily be restricted to variable regions, and those for strains from different genera or families need not be restricted to conserved regions.

Evidence of discordant variability in 16S sequences is also observed in taxonomic studies, where different variable region sequence analyses provide different taxonomic levels of classification for different bacteria (6, 29, 30, 34, 44). Computational tools used for sequence classification (13, 14, 45) do not weight nucleotide identities based on their evolutionary conservation across references. Instead, each nucleotide difference relative to a reference is weighted evenly, and as such an assumption forms that each nucleotide variation is similarly consequential (or inconsequential). Therefore, species-level classification is difficult when multiple intra-genus references have highly similar sequences. A single nucleotide divergence from the genus 16S consensus would be undervalued due to overall high similarity of the sequence with references, even if the change is indicative of a specific species.

Conserved rRNA nucleotides in a population are often those that form important interactions in ribosomes, whereas variable nucleotides are often those that form less important interactions (54). However, the evolutionary conservation of 16S rRNA nucleotides in species is often unexplored when investigating rRNA structures, especially in so-called variable regions where nucleotide variations are assumed to be frequent. Computational tools used to assess RNA
sequence secondary structures based on models showed that covarying mutations are important in order for 16S rRNA to maintain structure (55–58). Variable regions encompass helices of 16S rRNA, and therefore nucleotide changes in stems of such regions may require a covariation in order for 16S rRNA to form appropriate structures. However, computational tools have investigated rRNA secondary structures from few organisms, of which disparate 16S rRNA sequences such as those for *Escherichia coli* and *Clostridioides difficile* have been estimated to have different models (54). Model effects from a few nucleotide changes, as seen among species in a genus or among closely related genera are unknown.

rRNA secondary structures do not account for interactions with small subunit proteins and other translation factors that occur *in vivo*. Nucleotide changes may affect direct interactions or affect secondary structure formation which indirectly affect interactions with ribosomal components. Previous *in vivo* studies of *E. coli* 16S rRNA harboring single nucleotide mutations identified a few variable region mutations as “mildly influential” to *E. coli* growth (59). However, that study, as well as others, focused primarily on orthogonal expression effects brought about by *E. coli* 16S rRNA residue changes in conserved regions (59–61). Other studies evaluated *E. coli* ribosomes harboring non-*coli* 16S rRNA, which showed that growth rates (62, 63) and orthogonal translation activity (63) became reduced with lower 16S sequence identity (compared to *E. coli*). Although *Shigella* 16S gene sequences are highly similar to those of *E. coli*, ribosomes harboring their 16S rRNA also showed compromised translation activity (63). Put together, these studies suggested that 16S rRNAs with a few nucleotide differences, including single nucleotide variations in variable regions, can have biological impacts. However, no studies have evaluated the influences of variable region nucleotide variations associated with specific species.
Metagenomics Evaluation of Biofilms Formed on Polymers Submerged in the Ocean

Metagenomic sequencing can be a more efficient option than amplicon sequencing in cases where multiple genetic markers are analyzed, or the overall DNA present an environment sample must be analyzed. As such, several metagenomic sequencing studies have achieved identification of microbial species and their genetic components in soil, water, and sewages samples (64–66).

Due to ever increasing plastic pollution and their potential harmful effects, the interest of researching organisms that can degrade plastics have increased (67–70). Studies have reported species belonging to genera Bacillus, Pseudomonas, Vibrio, Sphingomonas, Streptomyces, and Burkholderia among bacteria, and Aspergillus, Penicillium, Candida, and Fusarium fungi. are capable of degrading polyethylene (PE), polypropylene (PP) (67, 68, 71–73). Alkane hydroxylase and laccase are the key enzymes involved in PE degradation (74) while the degradation enzymes for PP degradation are unknown (75). Biodegradation of polystyrene (PS), polyvinyl chloride (PVC), and polyurethane (PU) typically involve strains of Bacillus, Pseudomonas, Rhodococcus and Candida, which utilize esterases, lipases, and cutinases as enzymes responsible for degradation (74, 76). Previous metagenomic sequencing studies have identified polymer-degrading enzymes present in ocean environments (77–79), of which one study reported the potential for much richer diversity in polymer-degrading organisms in benthic environments (18). However, traditional tools for functional genomics do not provide a clear indication whether the enzymes identified are polymer-degrading (80–82). A recent study published a database sequences of the proteins associated with polymer degradation (83), which can hence be utilized to assess the presence of these enzyme genes in metagenomic sequencing studies.
CHAPTER THREE: METHODOLOGY

The methodologies illustrated in this chapter are separated into three major subheadings (underlined) based on the three studies conducted in this dissertation.

Beer Cultures and Microbial DNA Analysis

The methods illustrated in this subheading were derived from the manuscript titled “Microbial Communities in Retail Draft Beers and the Biofilms They Produce” published by the author of this dissertation, and who owns copyright over the manuscript (10) (Copyright information provided in Appendix A).

Beer Sample Collection and Culturing

Samples were collected from retail drafts that were supplied from kegs maintained at 4°C. The drafts were supplied using 1/2-inch clear PVC tubing, which exited the cooler and ran 25 ft ambiently to refrigerated draft service heads. The tubing was approximately 10 years old and exhibited turbidity on the interior surfaces. This retail location does not flush beer lines prior to daily service; therefore, the beers were collected as the first draws of the day without line flushing. Approximately 200 mL of draft samples of four beers from different breweries were collected into sterile cups: an American lager (“L,” unpasteurized, 4.5% ABV, ~15 IBU), an India Pale Ale (“I,” unpasteurized, 9% ABV, ~90 IBU) a hefeweizen (“H,” unpasteurized, 5.3% ABV, ~18 IBU), and an extra pale (“E,” unpasteurized, 5.7% ABV, ~40 IBU). These values were taken from the manufacturer websites, which are not referenced to maintain anonymity. After mixing by swirling, 50 mL was transferred to sterile conical tubes and placed on ice. Approximately 45 min later, each tube was centrifuged for 30 min at 3,500 RCF at 4°C and 45
mL of the cleared supernatants was removed by aspiration. The pellets were resuspended in the residual 5 mL, creating 10X “starter” stocks. 1 mL aliquots of each sample were then prepared: one was frozen at -80°C and the others were kept on ice to serve as inoculation sources that day. This process was repeated the following year from the same taps.

Solid substrates for biofilm development were formed by hand-stamp punching 0.25-inch plugs from 3/8 inch flexible PVC (Vinyl-Flex NFS-61, Advanced Technology Products, Milford Center, Ohio) and collected into a clean glass beaker. In a sterile laminar flow hood, the plugs were then submerged in a peracetic acid sterilizing solution (SporGon, Decon Labs Inc., King of Prussia, PA) for 3 hr, and then rinsed four times with 0.22 µm filter-sterilized HPLC-grade H2O. After the last wash, the residual water was drained, and each plug was transferred using sterile tweezers into sterile 1.7 mL microfuge tubes and stored until use.

Sterile growth medium was prepared by 0.22 µm filtering a commercial canned lager that was the same as the sampled American lager, except it had been pasteurized prior to canning. 1 mL aliquots were then aseptically transferred to the tubes containing the PVC pellets. A set of 5 uninoculated tubes was set aside as sterility controls for each year's study and triplicate experimental tubes were inoculated with 10 µL of the 10X microbe starter stock and mixed by vortexing. The culture tubes were then placed in a 20°C incubator for 2 weeks with an additional vortex mixing event after the first week.

**DNA Extraction from Starter Stocks and Beer Cultures**

Culture tubes were vortexed briefly to resuspend settled cells and a 500 µL aliquot was set aside as the planktonic cell fractions. The PVC plugs were then transferred with sterile
tweezers to a microfuge tube containing 1 mL of 0.22 µm filter-sterilized HPLC-grade H2O and vortexed to remove nonadherent cells. This washing step was repeated two more times.

Culture tubes were vortexed briefly to resuspend settled cells and a 500 µL aliquot was set aside as the planktonic cell fractions. The PVC plugs were then transferred with sterile tweezers to a microfuge tube containing 1 mL of 0.22 µm filter-sterilized HPLC-grade H2O and vortexed to remove nonadherent cells. This washing step was repeated two more times.

Bead ablation tubes were prepared by adding 100 µL of 0.1 mm zirconia beads (Research Products International, Mount Prospect, IL) to sterile 2 mL screw cap microcentrifuge tubes. 500 mL of a denaturing “extraction buffer” (5.5 M guanidinium thiocyanate, 100 mM potassium acetate, pH 5.5) was added to the beads before adding 200 µL of either planktonic cells or 200 µL of sterile HPLC-grade water and a PVC plug. The tubes were then agitated twice using a FastPrep-24 lysis 5G instrument (MP Biomedicals, Irvine, CA) using the “Escherichia coli cells” setting. After disruption, cell debris and beads were collected by centrifugation at 14,000 RCF for 5 min and 500 µL of each cleared supernatant was transferred to a clean tube. 200 µL of isopropanol was added and mixed by vortexing and the solution transferred to a DNA binding silica spin column (EconoSpin, Epoch Life Science, Missouri City, TX). After passing the solution through the column twice, the column was washed with 300 µL of extraction buffer, followed by three washes with “column wash buffer” (80% ethanol, 10 mM Tris-Cl, 0.1 mM EDTA, pH 8.0). The columns were then dried by centrifugation and the samples eluted in 50 µL of “DNA buffer” (5 mM Tris-Cl, 0.1 mM EDTA, pH 8.0). DNA samples were stored at -20°C.
DNA Library Preparation and Sequencing

Bacterial 16S V3-V4 regions were amplified by PCR using universal primers derived from S-D-Bact-0341-b-S-17 and S-D-Bact-0785-a-A-21 (38) containing unique adapters for the Nextera XT indexing kit according to the manufacturer's instructions (Illumina Inc., San Diego, CA). Fungal ITS2 regions were amplified using separate adapter primers derived from IST3_KYO1 and IST4_KYO1 (84). After adding unique indices, each sample's DNA concentration was determined using the Quant-iT PicoGreen kit (Molecular Probes, Inc., Eugene, OR) and equal mass portions of each were pooled prior to sequencing.

The amplicon pools were paired-end sequenced (250 cycles) using the MiSeq platform (Illumina Inc.) at either the Interdisciplinary Center for Biotechnology Research (year 1 data set, University of Florida, Gainesville, FL) or the UCF Genomics & Bioinformatics Cluster (year 2 data set, University of Central Florida, Orlando, FL).

Sequence Processing and Classification of Microbial Organisms in Beer

Trimmomatic was used to remove primer sequences and to filter out reads less than 150 bases long as well as reads with Phred quality (Q) scores less than 25 using a sliding window of 4 bases (85). VSEARCH was used to merge the forward and reverse sequences to generate the complete V3-V4 regions with a minimum total merged sequence length of 200 bases, minimum overlap length of 50 bases, and a maximum of 5 allowed mismatches across the alignment (86). Following merging, sequences with an expected error rate >1 were discarded. The remaining sequences were then dereplicated while counting the number of each unique sequence.

VSEARCH was used to denoise reads, remove chimeras, and to generate zOTUs. Parameters for zOTU clustering included occurrence of a minimum of 50 unique reads within a
100% identity threshold. All error-filtered merged sequences were then mapped to zOTU sequences based on 100% alignment to generate a table containing sequence read counts per zOTU for each sample.

Taxonomic classification of zOTU sequences was carried out using the Bayesian Lowest Common Ancestry (BLCA) software, which employs full-length query-hit alignment scores to generate a weighted probability for taxonomic allocation (87). For bacterial classification, the NCBI 16S rRNA database was used (updated 06/24/2021) and for fungal classification the UNITE v6 database was configured for usage with BLCA (88). Classification was declared based on the lowest common ancestor with a cumulative posterior probability ≥80%. Fungal zOTUs were additionally manually compared to the NCBI's nonredundant eukaryotic database using BLAST (89, 90).

The percent abundance of each zOTU was determined in each sample's data set and used to calculate ratios relative to a common reference (zOTU1, which was present in all bacterial samples) as previously described (91). Finally, the log\textsubscript{2} transforms of these reference frames were used to evaluate relative changes in the bacterial communities during culturing and to compare the relative abundances in biofilms and planktonic samples.

**Taxonomy Bush and Graphics**

Taxonomy bushes were generated using the Interactive Tree of Life (iTOL) (92). Data were sorted and processed using Excel (Microsoft, Redmond, WA). Data were plotted using Prism (GraphPad, San Diego, CA) and figures were generated using Illustrator (Adobe Inc., San Jose, CA). Processed zOTU sequences and data analytics are components of the associated Supporting Information.
Variable Region rRNA Sequences Influence Ribosome Performance

The methods illustrated in this subheading are based on a submitted manuscript titled “Hypervariable rRNA sequences influence ribosome performance” written by the author of this dissertation.

Strains and Plasmids

The *E. coli* *rrsA* locus was PCR amplified from strain BW30270 (Yale Stock Center, CGSC# 7925) initially using a forward primer that annealed in the upstream *hemG* gene to focus the amplification to that *rrs* allele. A second PCR was templated by the initial amplicon that generated a segment lacking the *rrsA* promoters, but retaining the native 5’ and 3’ processing regions, including the 3’ *ileT* and *alaT* genes (corresponding to residues 4,035,375 to 4,037,386 of the NCBI reference sequence NC_000913.3). This amplicon was then enzymatically assembled using NEBuilder (New England Biolabs, Ipswich, MA) with a DNA plasmid fragment such that *rrsA* expression was under the control of a *P araBAD* promoter and a downstream *rrnB* terminator pair derived from plasmid pBAD24N (93). The completed plasmid also contained a p15A replication origin and a *bla* selection gene. V1 sequence tags and other mutations were introduced into the plasmid 16S gene using PCR-based site-directed mutagenesis (94), including the V1 region sequence tags. The experimental host strain was the *rrn- recA::frt* BW30270 derivative, NB5.

Experimental Culturing, Harvesting, and Lysis

Experimental cultures were grown at 37 ºC in MOPS EZ Rich Defined Medium (Teknova, Hollister, CA) (95) supplemented with 10 mM sodium bicarbonate, and either 0.2%
glycerol (plasmid-born 16S gene expression partially repressed) or 0.2% arabinose (plasmid-born 16S gene expression fully induced). Overnight starter cultures also contained 150 µg/mL ampicillin. On the day of harvests, overnight triplicate cultures were diluted 1:100 in 30 mL of MOPS EZ Rich Defined Medium. Cultures were harvested during exponential phase (absorbance 600nm = 0.2), by promptly transferring cultures from 37 °C to 50 mL conical tubes containing 20 mL crushed-ice wash buffer (25 mM HEPES Tris, 150 mM sodium chloride, and 20 mM magnesium acetate, 1 mM chloramphenicol, pH = 7.6) which halted translation and aided in keeping ribosomes bound to mRNA. Cells were immediately pelleted (12,500 RCF for 3 min at 4 °C). After supernatant removal, cells were resuspended in 450 µL of cold (4 °C) lysis buffer (25 mM HEPES Tris, 100 mM potassium glutamate, 20 mM magnesium acetate, 1 mM chloramphenicol, 14 mM β-mercaptoethanol, and 1X Protease Inhibitor Cocktail Set V from Millipore Sigma, pH=7.6) and transferred to a cold 1.7 mL screw-cap tube containing 100 µL 0.1 mm zirconia/silica beads (BioSpec Products, Bartlesville, OK). Lysis was carried out using an MPBio FastPrep-24 5G homogenizer (QuickPrep adapter, 10 m/s, 1 cycle, 30s). Cell debris and beads were pelleted (18,000 RCF for 10 minutes at 4 °C), and the supernatant extracted. Absorbance of the lysate supernatant was determined at 258 nm and normalized to the lowest value among replicates using lysis buffer.

**Cell Fractionation**

10-40% sucrose gradients were prepared (solvated in 25 mM HEPES Tris, 100 mM potassium glutamate, 20 mM magnesium acetate, 14 mM β-mercaptoethanol, and 0.5% Tween 20) in polyclear centrifuge tubes (Seton Scientific, Petaluma, CA) using a Gradient Master™ instrument (BioComp Instruments, Fredericton, NB) the day of culture harvest and kept at 4 °C
until lysates were prepared. 200 µL of normalized lysates were added to the tops of the gradients. Cellular components were separated in the gradients using a Beckman Coulter Optima L-90K ultracentrifuge with a SW41 Ti rotor at 35,000 RPM for 3.5 hrs at 2 ºC. The gradients were then fractionated using a BioComp Piston Gradient Fractionator™ with a BioRad 2110 Fraction Collector while absorbance was measured at 258 nm.

RNA Extraction

RNA was extracted from 360 µL of each gradient fraction of interest or from a mixture containing 50 µL of total lysate and 310 µL of RNA buffer (10 mM bis-Tris, 0.1 mM EDTA, pH = 6.5). Samples were mixed with 200 µL 3 M sodium chloride, 3 µL of 10 mg/mL linear polyacrylamide, and 600 µL of acidic phenol/chloroform/isoamyl alcohol (125:24:1, pH = 4.3). The aqueous phase was extracted and washed twice with 300 µL stabilized chloroform. The extracted aqueous phases were then mixed with 700 µL of isopropanol and incubated in ice for at least 30 minutes followed by precipitate harvesting by cold centrifugation. Without disturbing the pellets, the isopropanol layers were aspirated, and the pellets washed sequentially with 75% and 95% ethanol. The pellets were then dried and resuspended in 50 µL of TURBO™ DNase reaction mix (~2 U/µL) (Invitrogen™, Waltham, MA) with 1 µL Murine RNase Inhibitor (~0.8 U/µL) (New England Biolabs ®, Ipswich, MA). After 45 min incubation at 37 ºC, 310 µL of RNA buffer was added and the RNA was re-extracted with phenol/chloroform and resuspended in 50 µL of RNA buffer.
**RT-qPCR**

Complementary DNA was generated using the iScript™ Select cDNA Synthesis Kit (BioRad®, Hercules, CA) following the random primer mix protocol according to manufacturer instructions. Completed cDNA reactions were diluted tenfold using DNase-free water and used as templates for qPCR using the SsoFast™ EvaGreen® Supermix Kit (Bio-Rad®, Hercules, CA). Three equal volumes of this mix were dispensed per well in a 96-well plate as technical replicates and qPCR reactions were carried out on a CFX96™ Real-Time System (Bio-Rad) programmed for an initial denaturation 95 ºC for 30 s and cycling at 95 ºC for 5 s followed by 60 ºC for 10 s for 40 cycles). Raw qPCR fluorescence data were exported to an online tool for quantification using global fitting (96). Seed values for qPCR reactions were used to calculate plasmid vs. chromosomal 16S rRNA abundance ratios and scores. Replicate measurements were averaged, and their values and coefficients of variations were used to calculate abundance ratios and the resulting associated standard deviations.

**Retrieval of 16S Gene V1 and V3-V4 Sequences and Relative Entropy Analysis**

The NCBI Assembly database was used to download *E. coli, Escherichia*, and *Shigella* feature table text files for strains (97). Advanced filters for feature tables included “Latest RefSeq”, “Complete genome”, and “RefSeq has annotation”. A Python script was used to extract 16S rRNA gene locations by using keywords “16S ribosomal RNA” and “ribosomal RNA-16S”. Thereafter, using a Python script that incorporated the NCBI efetch tool, the “+” strand sequences of all 16S rRNA gene copies were obtained as a FASTA file. Multiple sequence alignments were carried out using MUSCLE (98). AliView (99) was then used to remove column gaps observed for all *E. coli* str. K-12 substr. MG1655 16S sequences across the
alignments (NCBI RefSeq# NC_000913.3). *E. coli* V1 region sequences were extracted based on an *E. coli* MG1655 reference (8) to determine V1 region nucleotide conservation. Also, the alignment section corresponding to V3-V4 segment (between the S-D-Bact-0341-b-S-17 and S-D-Bact-0785-a-A-21 primer binding sites (8, 38)). Column gaps observed for all *E. coli* str. K-12 substr. MG1655 16S sequences across the alignments were removed (NCBI Accession# NC_000913.3). *E. coli* V1 region sequences were then extracted based on an *E. coli* MG1655 reference (8) to determine V1 region nucleotide conservation. A Python script incorporating Kullback–Leibler divergence ($D_{KL}$) was then executed to determine the V3-V4 positional relative entropy for observing a residue across alleles in a strain compared to the entire population (Equation 1 seen below) (100, 101). The output file contained $D_{KL}$ per residue position for each strain in the respective populations of study.

\[
\text{Positional } D_{KL} = \sum_{i=1}^{b} P_i \log_2 \frac{P_i}{Q_i}
\]  

(1)

Here, $D_{KL}$ values are position-specific in the sequence alignment. $P_i$ is the frequency of the observed $i^{th}$ residue (A/C/G/T) at the $i^{th}$ V3-V4 position across alleles in a strain. $Q_i$ is the reference frequency of that residue at that V3-V4 position across all sequences in the population of study. $D_{KL}$ was used to identify variations specific for *E. coli* strains in the population of *E. coli*, non-coli *Escherichia* strains in the population of *Escherichia*, and *Shigella* strains in the population of *Escherichia* with *Shigella*. For nucleotides with 0 occurrence, $P_i \log_2 \frac{P_i}{Q_i}$ was evaluated as 0. A computed value of 0 implies that there is no positional difference in the residue between the strain and the overall population. A high $D_{KL}$ value is indicative of a different residue in multiple alleles of a strain at that position relative to the majority of the population. A cumulative sum of $D_{KL}$ (c$D_{KL}$) at a position determined if that variation occurred across
multiple strains. Designations for strains that contributed to high $cD_{KL}$ values were determined if they were a common species.

**V3 Sequence Determination for Gammaproteobacteria and Clostridioides difficile**

A premade multiple sequence alignment of non-redundant Gammaproteobacteria 16S rRNA sequences was retrieved from SILVA Ref NR 99 Release 138.1 (102). Jalview (103) was used to obtain the consensus residue percent for the V3 region (8) between residue 433-497 using *Escherichia coli* MG1655 as a reference. Separately, 16S gene sequences from *Clostridioides difficile* str. 630 (NCBI Accession# NZ_CP010905.2) and *Escherichia coli* str. K-12 substr. MG1655 (NCBI Accession# NC_000913.3) were obtained and aligned using Aliview. Alignment positions corresponding to residues 433-497 for the *E. coli* reference 16S gene were considered as the *C. difficile* V3 region. The Needleman-Wunsch tool on EMBOSS for pairwise sequence alignment (104) was used to compare the *E. coli* and *C. diff* V3 sequences.

**Structure Analysis**

Chimera X was used to view and evaluate *E. coli* ribosomes (105). All illustrations used PDB 4V9D which contains a ribosome with bound tRNA and was previously used to determine RNA-RNA interactions in *E. coli* 16S rRNA (106, 107). Hydrogen bonds were illustrated using default parameters, ignoring intra-residue bonds. Putative V1 and V3 region secondary structures were determined using RNAfold (108). The predicted tagged V1 secondary structure was similar to a biochemically established *E. coli* V1 secondary structure (107) with the exception of a possible difference in interaction between the residue at position 76 and the residue at position 93 or 94.
Polymer Cable Sample Processing

Cables Jacket Materials and their Retrieval from the Ocean

Cables used contained an outer jacket made of four different synthetic polymers: 1) Hytrel; a high density thermo polyester elastomer (TPE), 2) Kynar; a form of polyvinylidene fluoride (PVDF), 3) Marlex; a high-density polyethylene (HDPE), and 4) Pinnacle Copolymer; a form of polypropylene (PP). The diameter of the cables with jackets was ~ 5 mm. These cables were submerged approximately 1 mile off the coast of South Florida, at an approximate depth of 70 feet. One-meter lengths of the cable were retrieved at five timepoints at 3, 6, 10, 12, and 17 months starting from the time of submersion in July 2021. The cables were covered in ~250 mL of a liquid preservative meant to prevent organism growth and RNA degradation (~ 3.5 M ammonium sulfate, 25 mM sodium citrate, and ~37.5 mM disodium-EDTA). Each cable was coated in this preservative in cleaned individual air-tight containers.

Seawater Retrieval

Four replicates of seawater at the same depth of the cables were sampled at each timepoint. Each replicate was dispensed into separate sterile conical flasks containing ammonium sulfate, sodium citrate, and disodium EDTA such that their final concentrations in a total volume of 40 mL were the same as the liquid preservative for cables.

DNA extraction from Cable Samplings and Seawater

Three ~3.6 cm fragments of the cable were taken from three regions: 1) where biofilm was visible along an 11 cm length of the cable (Visible Growth or VG), 2) where little to no biofilm was visible along an 11 cm length of the cable (Low visible Growth or LG), and 3) a
region where cables were clustered together and surrounded by visible biofilm (Clustered visible Growth or CG). The three fragments from each region were placed in a 2 mL bead ablation screw-cap tube containing 100 µL of 0.1 mm silica beads, as well as 300 µL PowerBead Solution and 50 µL solution SL from the DNeasy® PowerLyzer® Microbial Kit (QIAGEN©, Hilden, Germany). The tubes were then agitated at 2800 RPM for 5 min using a PowerLyzer® 24 Homogenizer (QIAGEN©, Hilden, Germany). The cell debris was pelleted by centrifugation at 10,000 RCF for 1 min at room temperature. The supernatant was dispensed into a fresh tube, followed by addition of RNase A (final concentration ~2 ng/µL) and incubated at room temperature for 5 min. The steps that followed were according to instructions for the DNeasy® PowerLyzer® Microbial Kit.

The seawater samples were centrifuged at 5000 RCF for 30 min at 4°C. The supernatant was removed and the cells resuspended in 300 µL PowerBead Solution. The cell suspension was transferred to a bead ablation tube with 100 µL of 0.1 mm silica beads followed by the addition of solution SL (50 µL) from the DNeasy® PowerLyzer® Microbial Kit (QIAGEN©, Hilden, Germany). The steps that followed were identical to DNA extractions from cables.

**DNA Quantitation, Normalization, and Sequencing**

The concentrations of extracted DNA were quantified using the Quant-iT™ Picogreen ® dsDNA Assay Kit (Invitrogen™, Carlsbad, CA), applying standard curves for high and low range DNA concentration detection. Samples were then normalized to a final concentration of 36 ng/µl in 50 µL, which included genomic DNA from *Escherichia coli* for samples with low concentration. The normalized DNA concentrations were checked using a Qubit dsDNA HS Assay kit (Invitrogen™, Carlsbad, CA) as per quality control recommendations for second
generation sequencing, making adjustments in concentration with *E. coli* genomic DNA. The normalized DNA samples were then submitted to GENEWIZ© Azenta Life Sciences© (South Plainfield, NJ) for 2x150 paired-end sequencing using the HiSeq Platform (Illumina©).

*Bioinformatics: Sequence Processing*

Sequencing reads were first processed using BBMap to remove adapter sequences, sequences with Phred (Q) score less than 30, and *E. coli* sequences. Paired-end sequences were obtained using Trimmomatic (85).

*Bioinformatics: Genus Classification and Relative Abundance Estimation of Polymer-degrading Genera*

Processed forward and reverse sequences were then mapped to a nucleotide database containing full genome references of archaea, bacteria, fungi, human, plants, protozoa, viruses, and human from NCBI RefSeq using Kraken2 (109). Unmapped sequences were then mapped to a protein database of the same references using Kraken2, utilizing a translated search with six reading frames. Bracken was then used to estimate genus-specific abundances (110). Bracken reports retrieved from the nucleotide and protein database searches were combined to estimate the overall abundances of genus-classified reads. Beta diversity between water and cable samples was determined by Bray Curtis dissimilarity using genus-classified read abundances. Abundances of classified reads relative to the total processed reads were determined. Identities and read abundances of genera for putative polymer-degrading species were extracted using a Python script that checked the presence of genera for known polymer-degrading species based on a reference list from previous studies (73, 83, 111).
Contiguous sequences were built from processed reads using MEGAHIT (112), applying a minimum contiguous sequence length of 200 bases. The contiguous sequences were then compared to protein sequences for polymer-degrading enzymes obtained from PlasticDB (83) using a local version of BLASTX (89). A custom Python script was then used to isolate matches and their contiguous sequences that achieved, at minimum, the default match criteria used in the online Plastic DB database (percent identity ≥ 30%, and E-value ≤ 10^{-6}) (https://plasticdb.org/annotategene). Abundances of processed reads that mapped to contiguous sequences corresponding to putative polymer-degrading enzyme genes were determined using BBMap and relative abundances determined with respect to the total processed reads per sample.
CHAPTER FOUR: FINDINGS

The findings illustrated in this chapter are separated into three major subheadings (underlined) based on the three studies conducted in this dissertation.

Microbial Analysis of Retail Draft Beers

The results illustrated in this subheading were derived from the manuscript titled “Microbial Communities in Retail Draft Beers and the Biofilms They Produce” published by the author of this dissertation, and who owns copyright over the manuscript (10) (Copyright information provided in Appendix A).

Establishing a Test Platform for Beer Microbiota.

Four beer draft taps at a single retail location were selected for study that delivered a lager (L), an India pale ale (I), a hefeweizen (H), and an extra pale ale (E) (Figure 1). Aliquots of each draft sample were used to inoculate three replicate cultures. The selected growth medium was the same brand of lager drawn from the tap but sourced from a can to avoid prior microbial contamination. To ensure sterility, the growth beer was also filter sterilized prior to delivery into the culture tubes. To provide a surface that appropriately represented the beer service lines in this system, the experimental culture tubes also contained uniformly dimensioned plugs of the same type of polyvinyl chloride (PVC) that comprised the retail service lines. These plugs rested at an angle submerged in the culture medium to allow for uninhibited beer exposure and to allow non-biofilm settling microbes to drift to the tube bottoms. The plug surface area and liquid volume were the same for each replicate. Upon sealing, no additional atmospheric exposures occurred until harvesting.
After 1 week, the samples were mixed by vortexing to redistribute the microbes and they were incubated for an additional week. A set of 5 uninoculated medium controls for each study year exhibited no turbidity after incubation and 10 mL platings of each on malt agar medium yielded no colonies.

Figure 1: Beer sampling and biofilm development. Beer samples were collected from four draft taps serving a lager (L), an IPA (I), a hefeweizen (H), or an EPA (E) as the first draws of that day. The microbes in each were concentrated 10-fold using centrifugation to create starter cultures and sampled for DNA extractions. Each culture tube contained sterile lager beer as a growth medium and uniform plugs of draft line plastic prior to inoculation with a starter culture. These cultures were allowed to develop for 2 weeks before extracting DNA from the planktonic and biofilm cells. The DNA in each sample was processed to establish the abundance of different microbes and then compared. Approximately 1 year later, a second sampling was performed from the same taps (which were still serving the same beers) and the experiment was repeated.
Identifying Bacteria and Fungi

The samples were processed to recover total DNA and polymerase chain reactions (PCR) were used to amplify either the hypervariable V3-V4 regions of bacterial 16S ribosomal genes or fungal ITS2 regions between the 5.8S and 28S ribosomal genes (38, 84). These PCR amplicons were then barcoded, pooled, and sequenced using paired-end Illumina technology (113). We obtained ITS2 PCR amplicons from each of the year 1 starter samples but were not able to recover amplicons from the year 2 samples, which suggests fungi were in very low abundance. To evaluate the microbial communities, the DNA sequences were computationally processed to identify the source genera and species, along with their relative abundances in the samples. We were unable to obtain bacterial PCR amplicons from the canned lager growth medium, so any genomes present were below the limit of detection and did not substantially contribute to the sequence collections.

Most bacterial genomes contain multiple copies of ribosomal genes and the 16S V3-V4 regions within them may differ in a single organism (6, 51). Moreover, subspecies (strains) of bacteria frequently have the same V3-V4 regions as other members of the species (6, 114). Fungal genomes can have tens or thousands of ribosomal gene copies, which can also be variable within a given species or strain (115–117). Therefore, while the counts of a given sequence and putative species names are informative for comparing community cohorts, they are not directly correlated with cell numbers or strain diversity.

Microbial Diversity in the Starter Samples

In the samples collected from the draft taps in the first year, each beer had a different and very rich community structure. In these beers, we identified 164 different V3-V4 sequences that
were derived from 98 species of bacteria. Within these samples, we also identified 18 ITS2 fungal sequences and were able to confidently assign 16 species as the sources of them. To gain insight into the dominant community members in these groups, we identified those sequence reads that were present at greater than 1% of the total reads in any given starter sample. A comparison of the community structures revealed that the most dominant bacterial members in the year 1 collection varied substantially between each style of beer, with Acetobacter, Fructilactobacillus, or Serratia as major members (Figure 2A). Likewise, each starter beer exhibited a notably different fungal community composition (Figure 2A). A highly abundant Saccharomyces cerevisiae sequence (zero-radius operational taxonomic unit [zOTU1]) was present in the hefeweizen sample, which was anticipated because hefeweizens are not filtered prior to service and they are visibly turbid from the brewing yeast. Without additional sequence information on the genomes of these organisms, we are unable to determine if this organism was the same one observed among the other samples because this ITS2 sequence is found in many S. cerevisiae strains. The other species are considered to have been beer contaminants.

For the samples collected in the second year, we identified 143 unique bacterial sequences derived from 72 species, most of which were the same as those observed in the year 1 collection. However, the relative abundances of each were markedly different, with Acetobacter having dominated all starters (Figure 2B). These comparisons highlight an important conclusion from this study: although major community members were similar between the first- and second-year collections, the relative abundances of the bacteria changed dramatically between sampling events. Thus, retail draft line communities can be dynamic and there is no particular pattern of bacterial abundance that could predict which beer they came from.
Figure 2: Bacteria and fungi present in the starter samples. Bacterial V3-V4 and fungal ITS2 hypervariable regions were sequenced and cataloged as zero-radius operational taxonomic units (zOTUs). These sequences were then assigned to source organisms at the genus or species level. (A) Year 1 bacterial and fungi abundances in the starter samples. Pie-charts illustrate the relative read abundances for the indicated organisms. Sequences with read abundances less than 1% of the total were grouped as “others.” (B) Bacterial abundances in the year 2 starter samples. Fungal ITS2 PCR amplicons were not recovered from the year 2 starter samples.

Dominant Culturable Microbes

We filtered the sequence collection to identify those bacteria that were reproducibly abundant after culturing, which indicates they replicated well in the lager medium. This processing reduced the number of bacterial genera to 31 for the year 1 samples and to 12 for the
year 2 samples. These genera were then compared for their evolutionary relatedness within the eubacterial kingdom (Figure 3). During this analysis, we discovered that approximately half of these genera were predominant in either the biofilm or planktonic culture samples, suggesting that those bacteria had preferred biological niches in our growth experiments (Figure 3).

**Figure 3:** Diversity of bacterial genera and their preferential growth. Taxonomic bushes illustrate the evolutionary diversity of bacteria that reproducibly grew well in the lab cultures, plotted from kingdom to genus for each year. Bacteria with sequence read counts greater than 1% of the total reads in that year are marked with asterisks. Cultured bacteria that were detected predominantly in biofilms are colored blue and underlined; those detected predominantly as planktonic are colored red.

Only a few of the yeasts grew well in the culturing experiments and none of those outgrowths exhibited a significant bias between the biofilm and planktonic fractions. In the samples inoculated with the lager-derived microbes, the *S. cerevisiae* and *Wickerhamomyces anomalus* strains grew well in all three culture replicates. The *Fusicolla aquaeductuum* that
provided the most sequence reads in this starter sample was not detected in five of the cultured samples and there were only a few reads in a planktonic sample that likely remained from the inoculum without any culture growth. In the cultures inoculated with the IPA starter, *W. anomalus* again grew well along with *Candida metapsilosis*, which was not abundant in this starter. The *Candida sake* that was dominant in this starter did not grow. Of the five prominent yeasts in the hefeweizen starter, only the *S. cerevisiae* grew out in the cultures.

Interestingly, a *Brettanomyces* strain that was not detected in this starter was detected in all cultured samples. Its absence in the data for the starter was likely caused by the dominance of the *S. cerevisiae* sequence reads in that sample, but there was no correlation between the abundance of *S. cerevisiae* and *Brettanomyces* in the cultures. Finally, of the four dominant yeasts in the EPA starter, only the *S. cerevisiae* grew out in the cultures. These observations indicate that, unlike the recovered bacteria, there was a stark difference between the yeasts that were present in the starter and their capability to grow in the cultures. Overall, the bacterial and fungal community structures that were present in these four starters were not maintained in the lab cultures, even for the cultures grown using the same lager brand.

**Preferences for Biofilm or Planktonic Growth**

The conclusion that some bacteria preferred occupancy in either the biofilm or planktonic communities was derived from a relatively straightforward visual inspection of the read count data for each growth experiment. However, that analysis overlooks bacteria that were abundant in both communities, but whose relative proportions in the biofilm and planktonic communities differed significantly after culturing. Bacteria that show a propensity to grow better in a biofilm relative to other community members may be capable of dominating when biofilms are
reestablished after draft line cleaning. To identify bacteria that exhibited such a behavior in our cultures, we applied an analytical method that compares the abundance of a given sequence read relative to a reference sequence that was present in all samples (91). This analysis has an advantage in that it does not require counting the absolute numbers of microbes in a given sample, which is intractable in biofilm studies. We elected to use the *Acetobacter* sequence (zOTU1) as a reference because it was abundant and present in all data sets. We were also able to leverage the outcomes of the replicated cultures to reveal reproducible behaviors.

For this analysis, we first filtered the data sets to only include sequence reads that were present at greater than 0.1% compared to the reference sequence in each of the samples. This filtering strategy avoided a pitfall caused by low abundance sequence reads: small stochastic differences in read counts between replicates can be incorrectly perceived as very large changes in relative abundance. We then calculated how the relative abundance of a sequence read in the cultured planktonic or biofilm samples changed compared to its abundance in the starter (relative differentials) (91). Taking a log$_2$ of those values provides an easier interpretation of any 2-fold increase or decrease around zero (zero indicates no change). The relative differentials were averaged across the three replicates and comparisons were made between biofilm and planktonic residency. For example, of the 14 sequence sets that passed the abundance filter in the year 1 lager culture set (a starter, three biofilm, and three planktonic samples in each set), there were seven instances where the relative read abundances were significantly different between the biofilm and planktonic environments (Figure 4A and 4B). In this representation, a positive value for the difference between biofilm abundance and planktonic abundance, the delta, means that the bacterium was more prevalent in that biofilm community with respect to the reference.
Figure 4: Evaluating bacterial biofilm preferences.
The sequence read counts in each sample were used to calculate relative abundances with respect to a common reference sequence in each sample (zOTU1). Those ratios were then used to establish changes in their relative abundances in the incubated biofilm or planktonic communities compared to their abundances in the starter samples. (A) Bar plots of the log$_2$ transforms of the relative differentials for bacteria in the incubated lager cultures (a 2-fold change is one unit on the ordinate axis). Error bars indicate the standard deviations between the three culture replicates. The ‘delta’ is the difference between the biofilm change and the planktonic change. (B) Bar plots of the deltas, with negative values indicating a preference for the planktonic niche and positive values for the biofilms in the cultured samples. Delta values from pairs that had significant differences between the biofilm and planktonic groups are marked with asterisks (t test *P values < 0.05). (C) Five additional significant biofilm deltas observed among the other seven culturing experiments.
Strong additional support for this analytical approach came from an interesting discovery we made regarding the five dominant *Fructilactobacillus lindneri* sequences in the year 1 lager experiment (zOTUs 3,4,7,8, and 9), where each was significantly overrepresented in those biofilms by ~50% (Figure 4B). *F. lindneri* (NCBI RefSeq NZ_CP014872.1) has seven 16S genes and the read counts among the samples for those zOTUs had nearly consistent proportions in all samples of 2:2:1:1:1, respectively. This correlation suggests there was a dominant *F. lindneri* strain with two copies of a 16S gene containing the zOTU3 or zOTU4 sequence, and three 16S genes with each of the others. A similar phenomenon was observed with *Loigolactobacillus backii*, where its five 16S genes appeared in a 3:1:1 ratio (zOTUs 14, 18, and 19), indicating that three 16S sequences were identical and the remaining two were different. An *L. backii* representative genome (NCBI RefSeq NZ_CP014890.1) also shows a 3:1:1 ratio of V3-V4 region sequences. In contrast to *F. lindneri*, the log-ratio comparisons between culture and starter revealed that these *L. backii* sequences consistently had positive values (Figure 4A), although insignificant deltas between the biofilm and planktonic fractions (Figure 4B). These results overall indicate that the abundances of 16S genes from these bacteria were changing in consistent proportions regardless of their copy number, which would be a requirement if they were members of the same genome. Moreover, this type of analysis may be useful in future studies for teasing out the number of strains within samples that have similar or shared gene sequences.

We performed the filtering and delta analysis for the seven other culturing experiments (three other beers from year 1 and four from year 2) and identified five sequences that also exhibited significant deviations between the biofilm and planktonic samples (Figure 4C). Interestingly, the *F. lindneri* zOTU3 and zOTU4 sequences were also overrepresented in the year
2 lager biofilms (by ;30-fold) and the corresponding zOTUs 7, 8, and 9 were again present in the same ratios as was observed in the year 1 experiment. Unfortunately, the read counts of the latter three were too low in the starter sample to survive our filtering protocol. Nonetheless, this organism grew well in the lager, and it became proportionally more abundant in the biofilms.

Overall, the microbial communities in each draft line became restructured over the course of a year and the dominant members changed, which indicates that the chemistry of the beers themselves was not fully defining those community structures. These observations highlight an important value in 'field monitoring' microbial communities in different industrial or clinical settings because single snapshots of diversity in complex microbial communities neither reflect their history nor predict their futures.
Ribosome Abundance Changes from 16S rRNA Mutations

The findings illustrated in this subheading are based on a submitted manuscript titled “Hypervariable rRNA sequences influence ribosome performance” written by the author of this dissertation.

Development of a Tagged 16S rRNA Tracking System

We developed an *rrs* expression plasmid to evaluate 16S rRNA variants *in vivo* using an *Escherichia coli* having unaltered *rrs* genes. This strain lacked *recA* to avoid recombination between plasmid and chromosomal *rrs* alleles. To measure the relative abundance of variants present at very low levels in different cellular fractions, we elected to modify the sequence of the plasmid 16S gene such that the expressed 16S rRNA could be selectively detected among abundant chromosome-derived 16S rRNAs using RT-qPCR. The candidate position for the alteration was helix 6, which comprises the central portion of the V1 hypervariable region. This helix generates a spur that extends away from the body of the ribosome (Figure 5A), and previous 16S rRNA variants that supported *E. coli* growth showed little sequence conservation in most of this region (62) (Figure 5B-C). In addition, other groups have used the tip of helix 6 to host RNA insertions that were used for the affinity purification of fully functional small subunits (118–120). Unfortunately, because insertions are flanked by wild-type sequences, we could not selectively monitor rare plasmid-born variants using a similar approach. As an alternative, we used nucleotide conservation among *E. coli* V1 regions as a guide to modify the plasmid *rrs* such that the 16S rRNA it encodes could be selectively detected in the presence of excess wild-type by virtue of the 'tagged' V1 region (Appendix B). We refer to the otherwise unaltered version of this 'tagged' clone as the parental version, into which additional alterations were introduced.
**Figure 5:** Structure and sequence analysis of *E. coli* 16S rRNA V1 region

The V1 region of the *E. coli* 16S rRNA was chosen for sequence tagging. **A)** The V1 region (red) protrudes away from the *E. coli* ribosome (PDB: 4V9D) and does not interact with atoms in the 50S proteins (green), 23S rRNA (light gray), 30S proteins (yellow), and most other 16S rRNA (black). The V1 region nucleotides (highlighted in the enlarged image) only form hydrogen bonds with other nucleotides in the region to maintain the spur structure. **B)** V1 region sequences of 16S rRNA genes and their alleles across *E. coli* strains (12,876 sequences) were aligned and the percent of each nucleotide at each position plotted. **C)** The alignment of V1 region sequences of 16S rRNA gene alleles from *E. coli* str. K-12 substr. MG1655 (NCBI# NC_000913.3) are shown with nucleotide variants colored and in lower case.
In prior studies evaluating the functionality of 16S rRNA variants \textit{in vivo}, the mutant versions were highly expressed \cite{61, 121–125}. Having high proportions of experimental 16S rRNAs within ribosome pools was used as an exploitable feature to quantify relative abundance or to generate sufficient material for selective purification. However, overexpressing additional rRNA can impart substantial stress if the expressed version is toxic or if it competes for assembly and maturation resources. Our strains harboring plasmids that expressed either untagged 16S rRNA, V1-tagged 16S rRNA, or an unrelated non-translated RNA of the same length became sickened by full plasmid induction and exhibited notably reduced growth rates (Appendix C). Using RT-qPCR to evaluate the relative abundance of overexpressed 16S rRNA, we observed that it accumulated to high levels, compared to the abundance of chromosomally-encoded versions (Appendix D). Moreover, we observed that the overexpressed 16S rRNA accumulated at the 30S gradient position, which suggested that particle maturation had become compromised (Appendix D). However, when the plasmids were not induced, plasmid-encoded V1-tagged 16S rRNAs were present at \(~1/500\) the level of chromosomal versions (Appendix D). Uninduced cultures grew normally (Appendix C) and the low level of plasmid-born 16S rRNAs became evenly distributed among the translation pool (Appendix D). Therefore, we chose to evaluate 16S rRNA variants in cells that did not have the plasmids induced. For a given mutant evaluation, abundance scores were calculated from informative regions of sucrose gradients and compared to the scores obtained from parallel control cultures that had expressed the parental 16S rRNA. (Figure 6A-C)
Figure 6: Establishing the abundances of 16S rRNA variants
Modified 16S rRNAs were evaluated in an E. coli strain with intact rrn operons. A) The E. coli rrsA gene was cloned into a plasmid under the control of a tightly repressed P_{BAD} promoter. The cloned rrsA was modified in its variable 1 (V1) region to contain a unique tracking tag sequence that was detectable using RT-qPCR. Other mutations were subsequently introduced in this tagged V1 rrsA for abundance evaluations. B) Fractionation of cell lysates using sucrose gradients allowed for isolation of 16S rRNAs in various stages of small subunit assembly and translation. The regions of 30S, 70S, and polysome material collected in this study are indicated. C) RNA was extracted from gradient fractions and used to establish the abundance ratio of plasmid-born 16S relative to chromosome-born in the same fraction. An abundance score was then calculated by comparing the abundance ratio in a given fraction to that of the unfractionated lysate.
Distribution of Toxic 16S rRNA Mutants

To evaluate our test system and also to establish comparative abundance scores for 16S rRNAs that are overtly defective, we chose to monitor decoding center mutants. Prior studies have established that alteration of the decoding center residue G530, A1492, or A1493 leads to the formation of small subunits that are dominant lethal (61, 124–126), potentially defective in initiation (121), and despite being detected in polysomes at low levels, they were compromised in their ability to translate mRNA (61, 118, 119, 124, 126). We independently introduced G530C, A1492U, and A1493U mutations into the V1-tagged rrs gene and calculated abundance scores for the encoded 16S rRNAs present at the 30S, 70S and polysome gradient positions. Because these RNAs had been expressed at very low levels, they did not slow the growth of the cultures. Each decoding center mutant was present at a higher abundance than the controls in the 30S and 70S fractions, but at a lower abundance in the polysome pools (Figure 7).

Figure 7: Evaluating decoding center mutants
Abundance scores were determined for plasmid-born 16S rRNA with separate decoding center mutations A1492U (red), A1493U (blue), and G530C (grey). Scores for mutants were compared to that observed for the parental 16S (y-axis) for the 30S, 70S, and polysome lysate fractions. Error bars represent standard deviations for biological replicates (n=3). Comparative statistics based on student’s t-test. P values <0.05 (*), <0.01 (**), < 0.001 (***)

43
Although this assay does not establish mechanistic causes for aberrant particle distributions, 30S subunits defective in maturation or initiation are expected to over accumulate near the 30S gradient position. Over accumulation in the 70S position may indicate a failure of the ribosomes to properly interact with other translation factors. Whereas a reduced occupancy in the polysome pool may indicate poor initiation, slow translation, increased turnover, or some combination. Therefore, the observed distributions of these decoding center mutants in our assay were consistent with the reported defects at several stages of translation.

_Disparate V3 Region Sequences affect 16S rRNA Abundances in Ribosomes_

After confirming that our test system provided outputs that reflected the quality of alternative 16S rRNAs, we turned our focus to variable regions that do not contain residues directly involved in translation. 16S variable regions from different bacteria tend to have similar sizes that symmetrically cluster around a mean value (127). One curious exception to this pattern is found in the V3 region encoding helix 17 (H17), the lower stem of which is involved in the assembly of the small subunit in _E. coli_ (128–130). The central region of V3 is highly variable and some species have V3 versions lacking approximately 25 residues from the central portion, while other species in the same genus do not (e.g., members of _Clostridioides_ Figure 8A) (127). This observation suggests that the central portion of V3 (encoding the outer stem-loop of H17) is not important for ribosome activity. However, we discovered that the central region of V3 is highly conserved in the class Gammaproteobacteria (which includes _E. coli_) (Figure 8B). Moreover, individual 16S alleles within the genomes of comparative reference strains have identical V3 regions (_E. coli_ MG1655 vs. _Clostridioides difficile_ 630), suggesting they are not entirely free to drift (Appendix E). In the mature _E. coli_ ribosome, the tip of H17 interacts with
residues of the V2 region (Figure 8C) (106). Unfortunately, there is no available ribosome structure of a C. difficile ribosome for comparison. These striking contrasts in the architecture and conservation of the V3 region raised the question of its importance in ribosome activity.

As an extreme evaluation of the importance of V3, we tested the performance of an E. coli 16S rRNA harboring the highly disparate C. difficile V3 region. The abundance scores for this variant indicated that most of the RNAs were trapped during subunit assembly, with very little progressing into the 70S or polysome pools (Figure 8D). To interrogate the conserved V3 Gammaproteobacteria region, we evaluated a 16S version harboring a single SNP (A465U) that was predicted to alter the interaction between the H17 tip and the V2 region in the E. coli structure (Figure 8C). Remarkably, this 16S variant performed nearly as poorly as the version containing the short C. difficile V3 and much of it was retained in the 30S fraction (Figure 8D). These findings support the hypothesis that conservation in the central portions of some V3 regions is due to a role in small subunit assembly that cannot be easily evolved away from, and it raises the question of how an organism could dispense of such an important region in all of its rrs alleles. These data also highlight the potential for improving taxonomic classifications by focusing on the identities of individual residues that are hallmarks of particular groups of bacteria.
Figure 8: Identification and performance assessment of disparate V3 region variants

The central portions of V3 regions are generally not conserved and fall into two length categories. A) *Escherichia coli* and *Clostridioides difficile* V3 region secondary structures were computationally predicted using RNAfold (109). The *C. difficile* V3 encodes a shorter helix and is missing the outer stem-loop. At the tip of the *E. coli* hairpin is residue A465 (arrow). B) An analysis of residue consensus in V3 region sequences revealed that A465 is present in over 99% of organisms in the class Gammaproteobacteria (total of 162,325 sequences). C) In the *E. coli* ribosome, A465 is located at the tip of helix 17 (lime-green) and forms potential hydrogen bonds with G203 and C215 in the V2 region (dark grey). Yellow spheres are residues of small subunit proteins (image rendered from PDB 4V9D). D) Abundance scores for *E. coli* 16S rRNA with *C. diff* V3 (orange) and an A465U transversion (violet) were determined relative to the parent 16S. Error bars represent standard deviations for biological replicates (n=3). Comparative statistics based on student’s t-test. *P* value < 0.001(***).
Positional Relative Entropy Reveals Strain- and Species-specific Residue Variations

To evaluate the capability of SNPs to provide taxonomic information at a higher level than overall homology, we focused on interrogating the V3-V4 region because of its prominent use and extensive reference data sets from environmental and clinical studies. The *Escherichia* and *Shigella* V3-V4 regions typically have 1-4 residue differences among them, so they are of no utility during classifications based on homology. Prior studies evaluated the average conservation of 16S rRNA nucleotides at each position based on their frequencies across single representative 16S sequences per bacteria (a form of Shannon entropy) and reported large deviations (or “surprises”) for nucleotides that deviated from the representative or consensus sequence (6, 29). However, Shannon entropy estimations of each strain or species would not weight SNPs that occur among alleles in a genome relative to the overall population. *Escherichia* and *Shigella* strains typically have seven copies of the *rrs* gene, each of which may have varying residues. To gauge the evolutionary change observed in a strain relative to the population, we computed a relative entropy ($D_{KL}$) (100) to compare the residue frequency observed at a V3-V4 position across alleles within a strain to the frequency at that position for the overall population (Figure 9A). Therefore, a residue variation in a single allele of only one organism of the population indicates less evolution, yields a low $D_{KL}$ value, and is less strain-informative. The same residue variation in all seven alleles of only one strain in the population indicates greater evolution, yields a high $D_{KL}$ value, and is more strain-informative. By taking the cumulative sum of $D_{KL}$ ($cD_{KL}$) at a position across strains in a population, we determined if the variation was prevalent in many strains and potentially species-informative.
Relative entropy analysis can be used to identify strain- or species-specific residues among variable region sequences in a population. A) Schematic depictions of variable regions of a multi-copy gene found in two hypothetical genera, X and Y. In a sequenced cohort of genus X organisms (population A), invariant residues at positions 1 and 4 provide no sub-genus information and SNPs observed at positions 2 and 3 are not associated with a particular species or strain, so their identities provide no information at those taxonomic levels. In genus Y, the SNP at position 1 (relative to that of genus X) is a strong genus indicator. In population B, occasional SNPs at position 2 provide no information because they are observed in single alleles in strains non-specific to a species. In population C, an SNP at position 3 is a strong strain indicator because it is present in all alleles in their genomes. In population D, occasional SNPs at position 4 indicate the presence of that species but provide no strain information. B) A Venn diagram illustrating the process of using positional relative entropy (\( D_{KL} \)) to identify informative V3-V4 residues, starting from inside to out.
We obtained V3-V4 region sequences from genomes of strains in the species *Escherichia coli*, the genus *Escherichia*, and the collective genera *Escherichia* and *Shigella*. We then computed the positional relative entropy of i) strains within the species *E. coli*, ii) non-*coli* strains within the genus *Escherichia*, and iii) *Shigella* strains within the total population of *Escherichia* and *Shigella* (**Figure 9B**). Our use of relative entropy therefore identified residue variations that discriminated certain species within the closely-related genera *Escherichia* and *Shigella*.

**Identification of Strain- and Species-informative rRNA Variants**

To illustrate the utility of positional relative entropy ($D_{KL}$), we used them to identify V3-V4 residues within *Escherichia* and *Shigella* that were strain- or species-informative. $D_{KL}$ values were determined by comparing the observed residue frequencies at each V3-V4 position per *E. coli* strain relative to frequencies at that position across 12,876 V3-V4 sequences from 1,850 *Escherichia coli* strains. We then plotted the maximum $D_{KL}$ values as well as the $cD_{KL}$ values per position, which aids in visually identifying strain- and species-informative variations, respectively. As such, two SNPs exhibited prominent entropy values across multiple *E. coli* strains (**Figure 10A**); an ‘A’ nucleotide at position 474 instead of the consensus ‘G’ (G474A) and an absence of ‘G’ at position 666 (G666-). To determine the relative weights of those values, we compared them to theoretical values of $D_{KL}$ and $cD_{KL}$ for hypothetical strains harboring 0 to 7 alleles with a specific SNP (**Figure 10B**). As the number of alleles with that SNP increases in a genome, so does the $D_{KL}$ value, reflecting an increasing uniqueness to that strain in the population. As the number of strains harboring that number of alleles increases in the overall population, the uniqueness decreases and that SNP becomes less strain-informative. For $cD_{KL}$
values, harboring more alleles again increases the information value; however, in this case, the more prevalent an SNP is in the total population, the more species-informative it becomes.

$D_{KL}$ and $cD_{KL}$ was determined for V3-V4 residue positions comparing non-coli Escherichia strains with all Escherichia strains. (13,429 V3-V4 sequences from 1,929 Escherichia strains). Two sets of high $D_{KL}$ and $cD_{KL}$ values were observed corresponding to TG591GA and CA647TC (Figure 10D), suggesting they are indicative of a strain and prevalent across strains. These polymorphisms were found exclusively in Escherichia albertii strains (total of 20) in multiple alleles in their genomes and were therefore strongly indicative of this species in a population of Escherichia.

Lastly, $D_{KL}$ was determined for Shigella strains, considering the positional frequencies of residues in the total population of Escherichia and Shigella as references (1,964 strains). This analysis revealed two prominent SNPs with high $D_{KL}$ and $cD_{KL}$ values, C488T and G748A (Figure 10E). C488T was found in most gene copies among Shigella boydii strains (total of 11) but was also in one or two gene copies among a few E. coli strains. G748A was exclusively found in Shigella dysenteriae strains (total of 10) and was always present in all 7 copies of the rrs gene. Also, low $D_{KL}$ values for G474A and G666- in the non-E. coli population sets indicates that these mutations were primarily present in the larger E. coli population. Overall, positional relative entropy revealed species-informative V3-V4 residues for E. coli, E. albertii, S. boydii, and S. dysenteriae.
Figure 10: Informative V3-V4 sequence polymorphisms among Escherichia and Shigella

Relative entropy was used to identify strain- and species-informative SNPs within Escherichia and Shigella 16S V3-V4 sequences. A) $D_{KL}$ peaks (red) correspond to SNPs at that position that are highly correlated with a particular strain among $E. coli$ strains. Peaks in a cumulative $D_{KL}$ ($cD_{KL}$) plot (grey) indicate an SNP that is prevalent across strains and may be specific to $E. coli$ species. For the evaluated $E. coli$ population (1850 strains), two positions showed high species $cD_{KL}$, corresponding to G474A ($cD_{KL} = 296.48$) and G666- ($cD_{KL} = 163.86$). B) Theoretical values for $D_{KL}$ and $cD_{KL}$ were calculated for up to 50 strains in the $E. coli$ population having an SNP in 0-7 out of 7 alleles. The $D_{KL}$ and $cD_{KL}$ values of the notable SNPs discussed in panel A are indicated for reference. C) $D_{KL}$ values were determined for non-coli Escherichia strains. Two polymorphisms, TG591GA and CA647TC, had high strain and species values and were found only in $E. albertii$ strains. D) $D_{KL}$ was calculated to identify Shigella strain- and species-SNPs within the large population of Escherichia and Shigella. C488T and G748A had high strain and species values and were $S. boydii$- and $S. dysenteriae$-specific, respectively.

Escherichia and Shigella V3-V4 Allele Variants Influence Ribosome Performance

A prior study showed that translation activity was compromised for Shigella 16S rRNA in $E. coli$ (63). We evaluated abundance scores for 16S rRNA with Escherichia and Shigella species-informative V3-V4 variations including variants observed in $E. coli$ (G474A and G666-),
*E. albertii* (TG591GA and CA647TC), *S. boydii* (C488T), and *S. dysenteriae* (G748A). Changes in abundance scores relative to the parent 16S illustrate that particular variable region changes can cause phenotypic effects.

*E. coli* ribosome structural studies showed that G474 is bonded with U458 ([Figure 11A](#fig11a)) and G666 is a is bonded with U740 ([Figure 11B](#fig11b)). Curiously, the G474A variant showed abundance scores higher than WT for all fractions, while G666- showed an increase in the 30S score, a decrease in 70S, and similar polysome scores as WT. These results indicate that the common *E. coli* V3 and V4 variants in 16S rRNA can perturb ribosome performances.

*E. albertii* V3-V4 polymorphisms occurred at two sites tandemly (TG591GA and CA647TC). In the *E. coli* ribosome, these residues form complementary base pairing ([Figure 11C](#fig11c)). Observably, these *E. albertii* variations also complement one another and form no other interactions. We evaluated scores with a single site change (UG591GA) as well as both sites changed (UG591GA and CA647UC). Both polymorphisms exhibited lower 30S scores, however only the single-site change exhibited lower scores in 70S and polysome fractions as well ([Figure 11D](#fig11d)). Altogether, these results indicate that 16S rRNA abundances in translating ribosomes can be maintained so long as these residues co-vary. Because four nucleotide changes are required to convert an *E. coli* V3-V4 into this *E. albertii* version, and because an intermediate in this conversion yields lower 16S abundances among polysomes, these alterations represent a comparatively large evolutionary hurdle.
Figure 11: Structure of V3-V4 informative residues in E. coli and abundance scores for Escherichia and Shigella species variants
The residue positions for Escherichia and Shigella species-informative SNPs were assessed in an E. coli ribosome structure (PDB 4V9D). Abundance scores were evaluated for E. coli 16S rRNA harboring informative Escherichia and Shigella species-informative V3-V4 residues. V3 residues in ribosome structures and abundance scores associated with their mutation are colored lime-green, and those for the V4 region are colored dark-blue. Unmutated residues are colored dark-grey in structures. A) G474 hydrogen bonds with U458 and G666 with U740. B) The abundance scores for 16S rRNAs harboring the species-informative Escherichia coli variations (G474A and G666-) were evaluated *in vivo*. C) Residues at sites for E. albertii-specific polymorphisms (UG591GA and CA647UC) complemented each other. D) Abundance scores for E. coli 16S rRNAs harboring E. albertii V3-V4 variant at one or both sites were evaluated. E) C488 and G748 are positioned to interact with G446 and C658 respectively. F) Abundance scores for E. coli 16S rRNAs harboring C488U (Shigella boydii SNP) or G748A (Shigella dysenteriae SNP) were evaluated. Error bars represent standard deviations for biological replicates (n=3). Comparative statistics based on student’s t-test. P values ≥ 0.05 (ns), <0.05 (*), <0.01 (**), < 0.001 (***)
Microbial Analysis of Polymer Cables Submerged in the Ocean

Establishing a Method to Extract DNA from Cable Biofilms and Seawater

Cables comprised jackets of four different materials (TPE, PVDF, HDPE, or PP) were retrieved from the ocean. Each cable retrieved from the ocean at a timepoint appeared to have varying densities of biofilm adhering to different regions. Because these different regions may have varying diversity of organisms, three regions were sampled per cable based on visibility of biofilms: 1) a continuous stretch of a single cable strand with visible biofilm covering the jacket (Visual Growth or VG), 2) a continuous stretch of a single cable strand where biofilms were less dense or not visible (Low Growth or LG), and 3) a region where cables were bunched with biofilms densely covering the cable cluster (Clustered Growth or CG) (Figure 12). Additionally, a single sampling of equivalent length was taken from each cable that was not submerged as comparisons. DNA was extracted from biofilms adhering to each cable sampling and from seawater that surrounded the cables at each timepoint. DNA concentrations were low for some LG samplings, the control unsubmerged cable, and seawater samples, and would not have cleared quality control recommendations for metagenomic sequencing. Therefore, *E. coli* genomic DNA was added to each sample to increase and normalize DNA extract concentrations for metagenomic sequencing (Figure 12).

The raw sequencing data obtained was first processed to remove low-quality sequences, as well as reads corresponding to *E. coli* spike-in based on comparison with a reference genome (NCBI# 0.000913.3). We note that downstream DNA abundance estimations for non-*E. coli* organisms may be minorly affected due to removal of sequences with high similarity compared to removed *E. coli* DNA, however, this effect would likely be much less for characteristic reads corresponding to organisms not closely-related to *E. coli*, such as polymer-degrading organisms.
Our DNA extraction and normalization method can hence serve as a reference for future studies requiring metagenomic evaluation for biofilm microbes.

**Figure 12:** Process for DNA extraction from ocean-submerged cables

Cables with four different types of jacket materials (TPE, PVDF, HDPE, or PP) were retrieved from a submersion site in the ocean. Clippings of the cables were sampled at three points based on cable clustering and density of biofilm observed: a single strand with thick visible biofilm (VG), a single strand with little-to-no visible biofilm (LG), and a cluster of cables surrounded by thick biofilm (CG). Each clipping was roughly 3.6 cm in order to fit into 1.5 mL screw cap tubes. The total length of clippings for each cable was 11 cm. Metagenomic DNA was extracted from samplings and their concentrations normalized with *E. coli* genomic DNA. Shotgun metagenomic sequencing was then performed, which yielded 150 base-pair fragmented sequences of the extracted DNA.

**DNA Read Classifications and their Relative Abundances for Biofilm on Unsubmerged Cables**

The processed sequences were compared to DNA in a custom annotated database that contained complete genomic information on various bacteria, archaea, fungi, plants, protozoa, viruses, and a few human genome references. The tool used for classification precomputes the ability of DNA fragments (called k-mers) from each reference to produce individualized
taxonomic identities and attaches an identifier to each fragment (109). The sequenced DNA fragments can then be compared to these individualized references and their identifiers directly called, making the classification process swifter. However, species classification may come with reduced accuracy (109), especially species are closely related and relevant references are not present in databases, which is a possibility for novel ocean-derived strains that were not previously isolated. Furthermore, previous studies typically isolated polymer-degrading strains from microplastics in soil or compost and while certain species may also be found in ocean environments, additional polymer-degrading species may also occur in the same genus. As such, sequences were conservatively classified to the genus-level.

The number of processed sequencing reads for unsubmerged cables were low (roughly 3600-4400), suggesting that few cells were present. The majority of these sequenced reads were successfully classified (80-88% relative abundance [RA]) of which human DNA was found to be dominant (50-58%) (Figure 13). Genera for other organisms detected above 1% RA were *Sphingomonas* (23-30%) and *Phyllobacterium* (1.2-1.6%).

Though the genera detected on unsubmerged cables can be considered as contaminants, we did not wish to rule out the possibility of detecting DNA corresponding to non-human organisms in biofilms in ocean waters at a depth of only ~ 70 feet. Furthermore, a previous study suggests that novel polymer degrading species may be detected in ocean waters (18) and species of *Sphingomonas* have previously been detected in marine waters (131, 132), some of which are capable of growing on synthetic polymers (69). Therefore, moving forward, only reads for human DNA were excluded because they were considered uninformative for this study.
Figure 13: Relative Abundances of DNA for Genera detected on Unsubmerged Cables
Read counts for classified sequences were determined relative to the overall processed read count for each unsubmerged cable. Bar graphs show relative abundances of separate classifications that were above 1%, and a combined relative abundance for those that were less than 1%. Relative abundances of reads that did not provide classifications are also shown (unclassified).

Relative Abundances of Classified DNA for Submerged Cables and Surrounding Water

Processed reads retrieved for each sample corresponding to submerged cable and water was classified and the abundances of classified DNA (RA) determined relative the total processed reads. The RA for classified genera were separated into those which have species known to degrade polymers and those which are not known to have polymer-degrading species (Figure 14). Classified DNA RA for submerged cables appeared to reduce with increasing submersion time, ranging from ~80% for low growth (LG) regions in PVDF and PP at the 3-month timepoint, to less than 15% for all cables at the 17-month timepoint. RA for genera with polymer-degrading species appeared to reflect the overall classified RA, with roughly 30% or less for LG regions at the 3-month timepoint to less than 1.5% at the 17-month timepoint (Figure 14). LG regions exhibited higher RA for classified DNA at the 3-month timepoint
compared to other sampling regions, however no discernible pattern for classified organisms were observed when comparing visible growth (VG) and clustered growth (CG) regions. Most processed reads for samples were not classified (likely due to the lack of genomic references in databases), yet DNA for putative polymer-degrading organisms, were less dominant compared to non-polymer-degrading organisms at each timepoint and each biofilm type.

The number of processed reads for each water sample replicate ranged from 978 to 82922, with an average of ~33000. To compensate for such low reads and determine overall genera content at each timepoint, reads for water replicates at each timepoint were pooled and abundances for DNA classified to the genera-level was determined relative to the total processed reads at each timepoint (Figure 14). The majority of sequences from water samples were classified for the 3-month, 12-month, and 17-month timepoints, however genera corresponding to putative polymer-degrading species were predominately only present at the 3-month timepoint.
Figure 14: Total Relative Abundances of Classified DNA for Submerged Cables and Water Samples

Relative abundances of DNA that classified to the genus-level were determined for each cable sampling. Classified genus abundances were split into those with known polymer-degrading species (based on a list obtained from literature reviews) and those that do not have polymer-degrading species. VG = Visible Growth on a single cable strand, LG = Low or no visible Growth on a single cable strand, CG = Clustering of cables surrounded by visible Growth.
Relative abundances (RA) for most genera were less than 5% across most samples, with the exception of *Sphingomonas* and *Moorena* species exhibiting RA ranging from ~5-20% only at the 3-month timepoint. Species for both genera have previously been detected in marine environments (131, 133) and *Moorena* species at the 3-month timepoint showed higher relative abundances than *Sphingomonas* at regions of visible growth (VG) and visible clustered growth (CG) than regions of low visible growth (LG). As such, *Moorena* species may have overtaken *Sphingomonas* in dominance, as more organisms settled on cables. However, the origin of *Sphingomonas* species on cables at the 3-month timepoint was uncertain, because *Sphingomonas* species were also detected as dominant organisms on unsubmerged cables as well as water samples at the 3-month timepoint.

Other genera that were detected above 1% RA were the eukaryote *Glycine*, as well as bacteria, *Streptomyces* and *Pseudomonas*. Detection of *Glycine* DNA appeared to be an oddity because they fall under the soybean family of plants, none of which are known to grow in marine environments. The same classification was also obtained when contiguous sequences built from the processed reads were classified. Several of such contiguous sequences (over 1000 bases) showed over 90% sequence identity to *Glycine max* (soybean) when searched using the online NCBI BLAST tool. Therefore, the DNA detected may be a genomic relative of *Glycine max*. Species of *Streptomyces* and *Pseudomonas* are known to occur in marine environments, some of which are known to degrade polymers (69, 134). Several other genera were detected between 1% and 0.1% RA, including *Vibrio, Bacillus*, and *Clostridium*, which are all known to have polymer-degrading species (68, 135, 136).

A beta diversity analysis was performed using Bray-Curtis distance estimation, which compares organisms and their abundances between samples and provides a score ranging from 0-
1, where 0 describes no difference between samples and 1 describes complete difference between samples. In some cases, samples from multiple cables appeared to cluster based on timepoints, but not based on specific cables nor specific regions on the cables (highlighted in red rectangles in Appendix F), suggesting that similar organisms settled on each cable at each timepoint. Notably, organisms and their abundances were diverse between cables and water samples as seen in previous studies (18, 68) (water samples shown as blue arrows in Appendix F). Because some genera known to have polymer-degrading species were identical across cables, their relative abundances were further assessed.

**A Focus on Genera for Putative Polymer-degrading Species on Cables**

A goal for this study was to evaluate the overall presence of polymer-degrading organisms on each cable. To do so, a list of genera for known polymer-degrading species was referenced (73, 83, 111) and read counts for those classifications extracted. The read counts for each genus was then summed across sampling regions at each timepoint for each cable and their abundances determined relative to their respective total processed reads.

Several genera were found to be above the overall average relative abundance (RA ~0.06%) for putative polymer-degrading genera and the RA for most genera reduced over time. *Sphingomonas*, which had the highest RA at 3-month timepoint, reduced by 40-300-fold by the 17-month timepoint. *Streptomyces, Pseudomonas, Vibrio, Burkholderia,* and *Rhodococcus* reduced 5-15-fold by the 17-month timepoint. *Candida,* and *Bacillus* showed similar RAs across all submerged cables, while *Acinetobacter* showed a two-fold increase in RA by the 17-month time point. By the last timepoint (17-months), *Candida, Pseudomonas, Bacillus,* and *Acinetobacter* were the dominant genera among those known to have polymer-degrading species.
Figure 15: Average Relative Abundance (RA) of Genera for Putative Polymer-degrading Organisms. Relative read abundances were averaged across cable samplings at each time point for each genus containing known polymer-degrading species. Genera are ranked from highest to lowest average relative abundance and those detected above the overall average (0.06%) are highlighted in the bar graph legend – the red box denotes those that were present on all cables, while those in the purple box were not present above the overall average on PVDF cable jackets.

Previous studies have shown that *Pseudomonas*, *Streptomyces*, *Vibrio*, *Rhodococcus*, and *Bacillus* species are known to degrade a variety of polymers, including PE and PP polymers in ocean environments (68, 78, 134, 135, 137). *Candida sp.* have been detected on PE, but not on PP (67). *Sphingomonas* species are primarily known to degrade poly-aromatic hydrocarbons (PAH) (138), however their species and related family members have been detected on
polyethylene (PE) and polypropylene (PP) in ocean environments (69). Therefore, the detection of several of these organisms on cables is expected. Interestingly, *Ralstonia*, *Ideonella*, and *Aspergillus* species were not as abundant as other genera, considering that they also have oceanic species known to degrade a variety of polymers, including PE (67, 76, 139). Notably, the top 19 genera (listed in Figure 15) were present on PE and PP cables above the overall average for at least one timepoint, but the top 18 and the top 12 were detected above the overall average for TPE and PVDF, respectively, showing that genera that settled on PE and PP jackets also settled on TPE and PVDF. Altogether, most genera for putative polymer-degrading species gradually reduced in dominance by the 17-month timepoint, further demonstrating that these organisms were less dominant compared to other organisms that settled on the cable. Furthermore, this study shows that genera that contain putative polymer-degrading species for a variety of polymers, can also settle on TPE and PVDF in a submerged ocean environment.

* A Focus on Gene Sequences that Corresponded to Polymer-degrading Enzymes

The Plastic DB database that was partially used to retrieve genera corresponding to putative polymer-degrading species contained protein sequences for enzymes responsible for polymer-degradation (140). Detection of these polymer-degrading enzymes in this study was determined using contiguous sequences built from processed reads (minimum contig length of 200 bases) followed by a translated search with protein sequences in the PlasticDB database. I did not restrict matches to only those known to degrade individual polymers used in this study because a) the enzymes known to be present on TPE, PVDF, and PP are largely unknown, and b) knowing the enzyme-content in this submersion site can serve as a reference for future biofilm studies using other polymers.
The total relative abundance (RA) of putative polymer-deteriorating enzymes was first determined for each cable sampling region at each timepoint. Unsubmerged cables showed RA for putative polymer-deteriorating enzymes less than 0.2%, the lowest of which was TPE (~0.005%) (Figure 16). The read depth of contiguous sequences for putative enzymes was less than 15 for all unsubmerged cables, suggesting that these enzymes were barely present. Among submerged cables, the RA of putative polymer-deteriorating enzymes for LG regions was either low or zero from the 6-month timepoint onwards, suggesting that organisms that contained these enzymes did not dominate in such regions (Figure 16). Conversely, VG or CG regions, which had dense biofilms, sometimes showed presence of enzymes at certain timepoints. While all cable samples showed less than 0.8% enzyme DNA RA, TPE showed even lower relative abundances (less than 0.045%) across most samples, except CG at the 10-month timepoint. Therefore, results suggest that TPE may be preferred over the other cable jacket materials to minimize presence of known polymer-deteriorating enzymes.
Figure 16: Total Relative Abundances of Putative Polymer-degrading Enzymes

The total read depth of contiguous sequences that corresponded to known polymer-degrading enzymes was determined relative to the overall processed reads for each cable sampling and at each timepoint. VG = Visible Growth on a single cable strand, LG = Low or no visible Growth on a single cable strand, CG = Clustering of cables surrounded by visible Growth, U = Unsubmerged cable.
The average RA of each putative polymer-degrading enzyme was then determined at each timepoint and for each cable. Several putative enzymes were at an RA above the overall average for submerged cables for at least one timepoint (~0.002%) (enzymes listed in Figure 17) however none were over 0.15%. Cutinases, PVA hydrolases, and PHB depolymerases were consistently above the overall average RA for all cables and at all timepoints. Esterases were approximately twofold higher than the overall average RA for PVDF, HDPE, and PP at all timepoints, but only above average for TPE at the 4-month timepoint. The other enzymes above the average RA varied for PVDF, HDPE, and PP cables at different timepoints but were often below average for TPE.

Notably, alkane hydroxylase and laccase, which are enzymes well known to degrade PE polymers (74, 75, 83, 141) were below average RA in HDPE samples, which is possible because HDPE is known to biodegrade poorly compared to LDPE (76, 141). Observance of cutinases, hydrolases, proteases, and esterases on several cables suggest the possible presence of polyurethane, potentially in the form microplastics that pollute the waters where the cables were submerged (74, 141). Interestingly, laccase enzymes were more often at a low relative DNA abundance even though PE plastics are the largest synthetic polymer pollutant in oceans (83). Perhaps this submersion site had fewer PE microplastics or that the microbes that possess laccase genes were present at low relative abundance on the highly durable cables. The observance of PVA hydrolase suggests the presence of polyvinyl plastics (75), which is the third largest synthetic polymer pollutant (83). The observance of PHB depolymerase suggests the presence of biopolymers such as PHB, PHA, or PPL, which are known to occur in ocean environments (71).
Relative read abundances were averaged for DNA corresponding to each polymer-degrading enzyme across cable samplings at each timepoint. Enzymes are ranked from highest to lowest average relative abundance and those detected above the overall average relative abundance (0.002%) are highlighted in the bar graph legend – the red box denotes those that were present on all cables, while those in the purple box were present on at least one cable, but not all.

The RA for putative polymer-degrading genera were not comparable with RA for putative polymer-degrading enzymes, suggesting that not all species within the genera possessed genes for polymer degradation or that polymer-degrading enzyme genes originated from organisms that were unclassified. Still, it is worth noting that species of *Pseudomonas, Bacillus,* and *Streptomyces* were present at higher-than-average RA and that they can have esterases, hydrolases, depolymerases, and cutinases which were dominant polymer-degrading enzymes (74, 76, 141). Unfortunately, without information on the structural integrity of cable jackets after
submersion the influence of polymer-degrading organisms and enzymes on cable jackets could not be determined. However, considering that an expectation of this study was to identify a cable with lower amounts of biofouling enzymes, all cables showed low relative abundances of these enzymes, but TPE more often showed the lowest of the four cables.
CHAPTER FIVE: CONCLUSION

This dissertation used case studies to evaluate the capabilities of microbial classification by amplicon sequencing as well as shotgun metagenomic sequencing; methods commonly used for DNA-based organism identification. In an amplicon sequencing study, I identified microbes that can inhabit retail draft beers and determined their ability to establish new communities in a controlled setting. Several bacteria known to spoil beers were found to occur on beer medium and as biofilms on PVC. These organisms included lactic acid bacteria (LAB) *Fructilactobacillus lindneri*, *Loigolactibacillus backii*, and *Levilactobacillus brevis*, as well as acetic acid bacteria *Acetobacter* (142–144). By classifying each unique variable region sequence (called zero-radius operational taxonomic units [zOTUs]) their relative abundance changes could be monitored across cultures without ambiguities, which showed that *Fructilactobacillus lindneri* bacteria preferred growth on PVC draft lines as biofilms over planktonic growth in beer medium. Furthermore, a recent investigation of the growth properties of *F. lindneri* showed that this bacterium requires low temperatures and anaerobic conditions for robust growth (143), suggesting that they may be missed by conventional aerobic culture-based monitoring. Put together, our findings showed that 16S gene variable region sequencing can lead to the detection of *F. lindneri* in spoiling beers, and if beer draft lines are not cleaned appropriately, *F. lindneri* colonies can persist and spoil fresh beers connected to the same line. Because zOTUs represented the actual variable region sequences of strains, we were also able to determine that *F. lindnerii* and *L. backii* sequences originated from respective individual strains due to their read abundances consistently occurring at ratios indicative of 16S alleles from their genomes. The dominant *Acetobacter* sequence in this study did not receive a species designation because it
was over 99% identical to 16S variable region sequences of *A. orleanensis, A. persici, and A. cerevisiae* (145). Species in this genus can consume ethanol, acetic acid, and lactic acid as energy sources in the presence of oxygen (146). Therefore, the dominance of *Acetobacter* in each of the year 2 samples suggests there may have been an alteration in the maintenance of the draft lines that allowed more oxygen exposure. Limiting oxygen exposure can reduce spoilage caused by *Acetobacter*. In summary, because spoilage microbes can be present on draft lines, I recommend developing a consistent guideline for line replacements as well as ensuring that equipment used for line cleaning is sterilized before use. Lastly, regular beer monitoring for spoilage microbes should be implemented, and this amplicon sequencing study can serve as a culture-independent method to do so.

Driven by the capability of 16S alleles to provide additional taxonomic information, I surveyed 16S gene variable region sequences in the closely related genera *Escherichia* and *Shigella* to determine if certain individual nucleotides can be informative of certain species. Though variable region sequences were over 99% identical among strains within these genera, using relative entropy calculations, I determined that pathogenic strains *Shigella boydii* and *Shigella dysenteriae* consistently possessed single nucleotide variants across genomic 16S alleles that distinguished them from the consensus variable region sequence for *Escherichia* and *Shigella*. Additionally, *Escherichia albertii* consistently showed a four-nucleotide variant across all copies of their 16S genes that also distinguished them from the genus 16S consensus. During taxonomic classification, allelic 16S sequences from individual strains is often ignored and each sequence is compared individually with references. However, the same nucleotide difference in all copies of the 16S gene (located at different positions in respective genomes) suggests that several additional non-random nucleotide variations have taken place. While this study
computationally evaluated variants within the commonly analyzed 16S variable regions 3 and 4, future studies can be conducted to determine species-informative variants in other variable regions as well as for any multi-allelic gene in any bacterial population. An in vivo study of *Escherichia coli* harboring the discovered species-informative polymorphisms for *S. boydii* and *S. dysenteriae* showed compromised ribosome quality, demonstrating that single nucleotide variable region mutations can have biological consequences. *E. coli* harboring half the covarying mutation from *E. albertii* exhibited compromised ribosomes, however when the full covariation was applied, ribosome quality was re-established in active translating pools. Therefore, a larger evolutionary hurdle was likely crossed for all four nucleotides to covary in *E. albertii* at all genomic alleles of their 16S gene. Put together, the few variable region sequence changes from the consensus genus can be indicative of certain intra-genus species, and when these nucleotide variants are placed in an intra-genus model strain, they can affect ribosome quality. Therefore, this study connects interpretations of 16S gene conservation with taxonomic, phylogenetic, and functional interpretations. In future studies, a computational weight may be applied for species classification when respective species-informative sequences are encountered in sequencing studies.

In our amplicon sequencing study, certain classifications (at the species, genus, or family level) were obtained at a lower confidence, which was likely due to the lack of suitable reference in databases or due to multiple close matches in the reference set. As such, reliance on the sequence of a single gene for computational classification can sometimes be problematic. Shotgun metagenomic sequencing can resolve these problems by classifying multiple sequences simultaneously. I surveyed organism communities on biofilms formed on cable jackets that were submerged in the ocean using metagenomic sequencing, focusing on putative polymer-degrading
microbes and enzymes. Though the number of sequences that could be classified progressively decreased with increased submersion time, those that were classified were at the genus-level. Results of the survey showed that relative abundances of genera for putative polymer-degrading species also reduced over time and the relative abundances of putative biofouling enzymes varied across timepoints but were consistently low. The cable jackets tested were composed of synthetic polymers made of TPE, HDPE, PVDF, or PP, which are known to have strong carbon-carbon bonds that prevent their deterioration from physical and biological influences (76, 141). Therefore, it is not surprising that relative abundances of DNA for putative polymer-degrading organisms and enzymes were low. The study did however conclude that certain bacterial and fungal genera for species that commonly adhere to a variety of polymers were present, such as *Pseudomonas, Streptomyces*, and *Bacillus*. Additionally, this study provided the first biofilm analysis of TPE and PVDF polymers.

Notably, I proposed improvements for species classification when analyzing ribosomal RNA genes, but I resorted to genus-level classification for metagenomics sequencing; a methodology that can often provide species-level classification (3). In both cases, computational classification of DNA sequences is dependent on the availability of suitable references in databases, and the lack of a suitable species reference can bias classification towards the next closest, but incorrect organism. In the case of the amplicon sequencing study, the species or genera classified were previously isolated and their genomes fully determined, and as such their ribosomal gene sequences could be closely matched in databases. However, the metagenomics study involved classifying microbes found in the ocean, several of which may not have previously been isolated, nor their genomes fully discovered. As such, to avoid potential species misidentifications, reads from shotgun metagenomic sequencing in the marine biofilm study
were conservatively classified to the genus-level from which further studies could be conducted to verify the species.

A common motivation derived from the studies conducted in this dissertation is the need to identify genetic sequences helpful for classification. For such improvements to occur, genomes of additional organisms need to be determined and compared to their related species. While genomic investigation of isolated organisms can provide a solution, an alternative culture-independent solution is to assemble genomes from metagenomic studies. Determination of metagenome assembled genomes (MAGs) have progressively shown success in identifying novel species (147, 148) and decrease the so-called “microbial dark matter” (149), however challenges still persist in avoiding contaminant sequences in genome assemblies (150). In the metagenome sequencing study in this dissertation, species read depth for each sample was insufficient to consistently obtain assembled genomes, however a future study may be pursued to obtain MAGs across samples. Even if a small percent of metagenome studies succeeds in providing MAGs, the number of species representative genomes still increases gradually which can help fill the gaps of unclassified sequence data. The confidence in establishing an MAG relies on strong correlation with an already established species (150), and therefore, we cycle back to the same motivation, which is the need to identify genes and genomes indicative of specific organism classifications. This dissertation partially resolves such a need by strongly correlating ribosomal RNA allelic sequences with certain species.
APPENDIX A: COPYRIGHT INFORMATION
FOR BEER LINE BIOFILM STUDY
APPENDIX B:
SPECIFICITY EVALUATION OF qPCR PRIMERS FOR TAGGED V1 16S
Specificity evaluations of qPCR primers for tagged V1 and WT V1: PCR primers that targeted the tagged V1 or WT V1 sequences were evaluated for selectivity using qPCR. A) Putative secondary structures of the WT V1 region (derived from \textit{rrsA}) and tagged V1 region bearing various sequence modifications are shown with qPCR primer-binding sites (red for primers targeting tagged V1 and green for primers targeting WT V1). Nucleotides of the tag sequences are in bold and capitalized. The MS2 tagged V1 is the WT V1 with an inserted sequence at the tip of the loop, which was previously used to purify small subunits. The swap-tagged V1 has a portion of the WT stem sequence of helix 6 swapped. The custom tagged V1 sequences contain modifications in the stem sequence that have predicted structures similar to WT V1. B) qPCR was performed using purified 16S genes containing the tagged or WT V1 sequences as templates and primers targeting respective tagged sequences or the WT sequence. Templates were normalized to have equal amounts. Specificity was determined by taking the ratio of qPCR seed values for the targeted 16S relative to the non-targeted version. For example, the specificity of the custom tagged V1-1 primers was determined from the ratio of qPCR seeds in reactions with the tagged V1 16S genes (targeted) relative to the WT V1 16S genes (non-targeted). The specificity of the WT V1 primers was determined from the ratio of qPCR seeds in reactions with the WT V1 16S genes (targeted) relative to each tagged V1 16S genes (non-targeted). qPCR primers targeting custom tagged V1-1 obtained the highest specificity and hence the plasmid \textit{rrsA} was tagged with this sequence.
APPENDIX C:

E. coli Culture Growth Rates for Different Plasmid Expressions (With Controls)
E. coli culture growth rates for different plasmid expressions (with controls). Doubling times of *E. coli* cultures during exponential phase were determined when plasmids expressed the tagged V1 16S rRNA (Parental 16S) either under uninduced (green) or induced (red) conditions, or when plasmids expressed the untagged WT 16S (orange) or a non-translated RNA fragment of the same length from a gene encoding a tail protein of phage lambda (*gph*) under induced condition (black). Comparative statistics represent student’s t-test results. *P* value < 0.001 (***).
APPENDIX D:
PERFORMANCE ASSESSMENT OF PARENTAL 16S
FOR VARIOUS EXPRESSION CONDITIONS
Performance assessment of parental 16S for various expression conditions. The plasmid-born 16S rRNA was detected by RT-qPCR in 30S, 70S, and polysome pools for uninduced and induced culture growth. A) The abundance scores of plasmid vs. chromosome-born 16S RNA was determined for RNA extracts from ribosome small subunit assembly (30S), ribosome formation (70S), and translating ribosomes (polysomes). B and C) RT-qPCR seed values were obtained for uninduced parental V1-tagged and WT chromosome-born 16S cDNAs in each gradient fraction and were normalized to their respective maxima (observed for fraction 6 containing 70S material, parental 16S cDNA seed = 6.6x10^8, WT 16S cDNA seed = 2.8x10^5). Error bars represent standard deviation for biological replicates (n=3).
APPENDIX E: *E. coli* AND *C. diff* V3 REGION ALIGNMENT
**E. coli and C. diff V3 region alignment.** A multiple sequence alignment using MUSCLE within Aliview was performed for 16S gene sequences from *Clostridioides difficile* str. 630 (NCBI# NZ_CP010905.2) and *Escherichia coli* str. K-12 substr. MG1655 (NCBI# NC_000913.3). The V3 region is highlighted in a black rectangle. Residues are colored as: A in green, C in blue, G in black, and T in red. Alleles are identical among genomic copies for the respective strains.
APPENDIX F:
BETA DIVERSITY COMPARING WATER AND CABLE SAMPLES
**Beta diversity analysis of cable and water samples.** Diversity between cable samples and water at each timepoint was determined by Bray-Curtis distance and represented as a Neighbor Joining Tree. The distance between two samples is determined by measuring horizontal branches that link to samples using the scale in the bottom left corner. Distances range from 0-1, where 1 means that organisms and their abundances are completely different (hence, “distant”). Red rectangles show samples that exhibited shorter distances in specific timepoints. Blue arrows show the water samples.
LIST OF REFERENCES


15. Schloss PD. 2021. Amplicon Sequence Variants Artificially Split Bacterial Genomes into Separate Clusters. mSphere.


