Selective Enrichment Through Capture (SEC): A Technology for Retrieving Specific DNA Sequences from Complex Mixtures

Kelley Gwin Nunez
Louisiana State University and Agricultural and Mechanical College

Follow this and additional works at: https://digitalcommons.lsu.edu/gradschool_dissertations

Recommended Citation
https://digitalcommons.lsu.edu/gradschool_dissertations/514

This Dissertation is brought to you for free and open access by the Graduate School at LSU Digital Commons. It has been accepted for inclusion in LSU Doctoral Dissertations by an authorized graduate school editor of LSU Digital Commons. For more information, please contact gradetd@lsu.edu.
SELECTIVE ENRICHMENT THROUGH CAPTURE (SEC): A TECHNOLOGY FOR RETREIVING SPECIFIC DNA SEQUENCES FROM COMPLEX MIXTURES

A Dissertation

Submitted to the Graduate Faculty of the Louisiana State University and Agricultural and Mechanical College in partial fulfillment of the requirements for the degree of Doctorate of Philosophy

in

The Department of Biological Sciences

by
Kelley G. Núñez
B.S., Louisiana State University, 2009
December 2015
ACKNOWLEDGEMENTS

I would first like to thank my advisor, Dr. John R. Battista who has guided me in becoming a well-rounded scientist. You have been an excellent mentor and have always been available for discussions. I have learned an enormous amount over the last five years about what it takes to be a scientist and I am extremely grateful for your patience, guidance, and freedom to pursue a wide variety of projects. I would also like to thank Dr. Gregg S. Pettis, Dr. Brent C. Christner, and Dr. James V. Moroney for allowing me to attend your laboratory meetings and your guidance on collaboration projects. I would also like to thank the rest of my graduate committee Dr. Mark Batzer, Dr. Charles Lee, and Dr. J. Cameron Thrash for your feedback and guidance. I would like to thank my former lab members, Kathir Selvam and Jana Duncan for their help during lab meetings. I would also like to thank the undergraduates in the lab, Dieu, Kirstie, Sarah Beth, Trent, and Meredith for allowing me to mentor you on microbiology and life. I want to thank Noelle Bryan who has become a close friend during my time in graduate school. I also want to thank members of Dr. Brent Christner’s laboratory, Amanda Achberger and Shawn Doyle for their helpful feedback and suggestions.

I want to thank Dr. J. Cameron Thrash and Dr. Jeremy Brown for their help with learning bash script, Unix commands, and using LSU’s high performance computing cluster. Part of my research would not be possible without the help of Dr. Jerome La Peyre from the School of Animal Sciences, Scott Herke in the Genomics Facility, and faculty at LSU’s Center for Computation and Technology.

I want to thank my parents for their support over the years and instilling in me the confidence to pursue my dreams. I want to thank my husband’s family for their support. I also want to thank my brother, sister-in-law, and sister for their encouragement and many trips to
Baton Rouge to visit me during stressful times. Finally, I want to sincerely thank my husband Kevin, and my son Keegan. Kevin’s support throughout the years has been critical to my success and I could not have done this without his love, dedication, and calming presence.

This work was supported by the United States Department of Justice, Office of Justice Programs (2010-DN-BX-K190) and National Aeronautics and Space Administration grant (NNX10AN07A).
# TABLE OF CONTENTS

ACKNOWLEDGEMENTS ........................................................................................................... ii

ABSTRACT ..................................................................................................................................... vi

CHAPTER 1. INTRODUCTION ............................................................................................... 1
  Streptavidin-Biotin Capture Techniques .................................................................................. 2
  Overview of Selected Enrichment through Capture (SEC) .................................................... 3
  Microbial Ecology – Determining Microbial Diversity .......................................................... 4
  Objectives of This Study ......................................................................................................... 8

CHAPTER 2. METHODS ......................................................................................................... 9
  SEC Capture Primers Design .................................................................................................. 9
  Linear Amplification Reactions ............................................................................................. 9
  Retrieving the Captured Fragments ....................................................................................... 12
  The Mock Community ......................................................................................................... 13
  Quantitative PCR (QPCR) Amplifications ............................................................................. 13
  PCR Amplifications ............................................................................................................. 14
  Illumina HiSeq 2000 Sequencing ......................................................................................... 16
  Trimming of Illumina Datasets ............................................................................................. 17
  Read Mapping of Datasets .................................................................................................... 18
  Generation of Simulated Datasets ......................................................................................... 18
  QIIME and EMIRGE Analyses ............................................................................................. 18
  Statistical Analyses ............................................................................................................. 20
  Oyster, Sediment, and Seawater Sample Collection .............................................................. 21
  DNA Extraction of Oysters, Sediment, and Seawater ............................................................ 22
  The C. reinhardtii Insertion Library ....................................................................................... 23
  Characterized C. reinhardtii Insertion Mutants ..................................................................... 24

CHAPTER 3. SELECTED ENRICHMENT THROUGH CAPTURE (SEC): METHOD FOR RETREIVING SINGLE LOCI FROM COMPLEX MIXTURES ......................................................... 25
  Introduction ......................................................................................................................... 25
  Results ................................................................................................................................. 27
  Discussion ........................................................................................................................... 46

CHAPTER 4. DETERMINATION OF THE MICROBIAL DIVERSITY WITHIN THE EASTERN OYSTER, CRASSOSTREA VIRGINIA BEFORE AND AFTER DEPURATION TREATMENT, AND COMPARISON WITH SURROUNDING SEDIMENT AND SEAWATER ................................................................................................................................. 54
  Introduction ......................................................................................................................... 54
  Results ................................................................................................................................. 58
  Discussion ........................................................................................................................... 71
ABSTRACT

The capacity to selectively target DNA sequences within a complex mixture is a useful feature for genomic studies. Several methods aimed at undertaking this feat have utilized biotinylated oligonucleotides and streptavidin beads to capture DNA but have low efficiencies, require PCR, or cannot be combined with high-throughput sequencing. The work presented here developed a protocol referred to as selected enrichment through capture (SEC) that uses a biotinylated “capture primer” to target and concentrate specific DNA sequences at high efficiencies, eliminates the use of PCR, and can be combined with high-throughput sequencing.

The effectiveness of SEC was evaluated in a series of studies that determined the efficiency, specificity, and recovery of specific sequences from DNA mixtures. Initially, it was established that SEC could retrieve a single locus as large as 6,200 bases from 3 ng DNA with high yield. This justified attempts to retrieve 16S rRNA genes from DNA from an artificial mock community and DNA extracted from the environment. Combining SEC with high-throughput sequencing allowed for the identification of all 20 species within an artificial community, reconstruction of full-length 16S rRNA genes, and predictions of relative abundances within an order of magnitude of reported values. SEC successful captured 16S rRNA genes from bacteria from seawater, sediment, and in the over abundances of eukaryotic DNA that facilitated the characterization of the microbiome of the Eastern oyster, C. virginica. SEC was further demonstrated to be successful in recovering 16S rRNA genes from the low biomass environment of glacial basal ice with 21 pg/μL of starting DNA. Additionally, SEC was successful in retrieving insertions in the eukaryote Chlamydomonas reinhardtii.

SEC has three advantages over other enrichment methods. 1) SEC does not rely on the PCR to enrich samples prior to sequencing. 2) Only a short segment of sequence adjacent to one
side of the targeted region needs to be known allowing for adjacent uncharacterized sequence to be recovered. 3) Sequences of interest can be isolated, concentrated, and analyzed without PCR amplification or interference from more abundant DNA found in the sample.
Retrieving specific DNA sequences from complex mixtures offers several advantages. One advantage is to remove targeted DNA from samples that contain contaminants like humic acids. By removing DNA of interest from these substances one can perform downstream analyses that include PCR (Jacobsen 1995). Another advantage for capturing DNA is the ability to concentrate a targeted sequence of interest. Capturing the DNA will remove the DNA of interest from a larger pool of DNA thereby concentrating the targeted sequence for downstream analysis. This can be particularly useful when targeting a gene present in small amounts (Jacobsen 1995). Lastly, another advantage in capturing specific DNA sequences is reducing the cost of reagents for downstream analysis. When performing high-throughput sequencing, capturing specific DNA sequences can reduce sequencing costs and reagents by avoiding the need to sequence entire genomes, instead sequencing only genes of interest (Mertes, Elsharawy et al. 2011).

In order for effective targeted capture of DNA, the method must be specific, efficient, require little input, and be cost effective. While PCR is a suitable method to accomplish this task, several issues limit the use of PCR. PCR requires the use of primers both up and downstream of the locus of interest. This is particularly important when PCR amplifying the 16S rRNA gene as primer sequences and primer pairs can greatly impact the specificity of amplification (Klindworth, Pruesse et al. 2013). Additionally, low abundant sequences may be underrepresented (Jumpstart Consortium Human Microbiome Project Data Generation Working 2012) and as a result hinder the efficiency of PCR. Moreover, samples that contain the contaminant humic acid would inhibit PCR (Jacobsen 1995). Bead-based methods that take
advantage of the strong bond between biotin and streptavidin have been used to target and retrieve specific DNA sequences with success.

**Streptavidin-Biotin Capture Techniques**

Several streptavidin-biotin capture methods have been developed in the last two decades (Tagle, Swaroop et al. 1993, Jacobsen 1995, Sterky, Holmberg et al. 1998, Briggs, Good et al. 2009, Gawronski, Wong et al. 2009). Tagle et al. (1993) used cosmids containing specific DNA sequences as bait. The cosmids were fragmented and linkers were ligated to the ends. cDNA libraries were then added to the biotinylated cosmid library and allowed to hybridize to the biotinylated cosmid fragments. The biotinylated cosmid-cDNA hybrids were then retrieved using streptavidin-coated magnetic beads. The cDNA was eluted using heat and underwent PCR amplification. Jacobsen (1995) used biotinylated primers to capture bacterial DNA as a way to remove the DNA from PCR inhibitory contaminants in environmental samples. Biotinylated probes specific to a single gene were used to saturate the surface of streptavidin-coated beads. Genomic DNA, isolated from soil containing inhibitory humic acids, was fragmented and combined with the biotinylated probe-coated beads. The gene of interest hybridizes to the beads and could be extracted from the humic acid. Sterky et al. (1998) used biotinylated primers to retrieve DNA sequences inserted into bacterial artificial chromosomes (BACs). A biotinylated primer designed to anneal to a known vector sequence within the BACs and the primer was extended into the cloned insert using a DNA polymerase (Sterky, Holmberg et al. 1998). The captured DNA had an average length of 1,000 bases (Sterky, Holmberg et al. 1998). Briggs et al. (2009) combined capture with biotinylated DNA fragments by streptavidin-coated beads with high-throughput sequencing, developing a method called primer extension capture (PEC) to retrieve degraded Neanderthal mitochondrial DNA (mtDNA). A biotinylated primer targeted a
specific DNA region within the mitochondrial genome, and a single linear amplification reaction was performed to extend the biotinylated primer from extensively degraded DNA (Briggs, Good et al. 2009). False priming with extension of the biotinylated primers did occur and multiple PEC reactions were necessary most likely due to low template DNA signal (Briggs, Good et al. 2009).

The methods described above relied either on hybridization of biotinylated oligonucleotides or extension of biotinylated primers. The drawbacks from each method include multiple subsequent rounds of PCR amplification to determine captured DNA sequence (Tagle, Swaroop et al. 1993, Sterky, Holmberg et al. 1998, Briggs, Good et al. 2009), low efficiency (0.004%) of captured DNA (Tagle, Swaroop et al. 1993), false capture of non-targeted DNA up to 85% (Tagle, Swaroop et al. 1993, Briggs, Good et al. 2009), and small captured DNA length (Sterky, Holmberg et al. 1998). The shortcomings of these methods were addressed in a new method that took sought out to improve on these limitations.

**Overview of Selected Enrichment through Capture (SEC)**

A method was developed that utilizes the strong bond between streptavidin and biotin and improves on previously developed streptavidin-biotin methods by providing higher efficiency, and longer captured DNA sequences. SEC is also intended to facilitate in the characterization of bacterial community structure without PCR amplification to avoid the potential biases observed when PCR is used. Like PCR-based amplifications, this method is designed to increase the concentration of 16S rDNA fragments for high-throughput sequencing, but it avoids PCR and relies instead on the complementarity between a biotinylated oligonucleotide and conserved sequences in the 16S rRNA gene to isolate the sequence. This biotinylated oligonucleotide not only provides specificity, but also acts as a primer for DNA
DNA isolated from an organism or an environmental sample is annealed to a DNA sequence of interest using a “capture primer”, labeled with a biotin molecule attached at the 5’ end (Figure 1.1). Linear amplification is then performed with Taq DNA polymerase which extends the biotinylated capture primer into the adjacent DNA sequence. The DNA is linearly amplified with the template number doubling for each round of amplification. Following sixty cycles of linear amplification, the entire reaction is added to streptavidin-coated magnetic beads. The biotinylated DNA then binds to the streptavidin on the surface of the magnetic beads that contains the capacity to bind 70 pmols of 4 kilobase biotinylated DNA (KilobaseBINDER™ kit, Life Technologies, Grand Island, NY). A magnet is used to separate the biotinylated DNA-streptavidin beads and the mixture is washed several times to remove the remaining reagents. The biotinylated DNA is then eluted using heat to break the streptavidin-biotin bond.

**Microbial Ecology – Determining Microbial Diversity**

Microbial ecology is the study of how microorganisms interact with one another and their environment. Many studies of microbial ecology have focused on prokaryotes, Bacteria and Archaea. Prokaryotes are ubiquitously distributed, being found in a wide range of environments from the upper atmosphere to the sub-seafloor (Schippers, Neretin et al. 2005, Smith, Timonen et al. 2013). Prokaryotes are diverse phylogenetically and physiologically, playing vital roles in the geochemical cycling of carbon, nitrogen, oxygen, and other important elements. Microbial ecology describes the diversity of the microorganisms present in an environment; an effort that is complicated by the fact that less than 1% of prokaryotes can be cultured as determined by differences in counts observed from direct plating on media versus direct microscopic cell counts (Staley and Konopka 1985).
Figure 1.1 Overview of Selected Enrichment through Capture (SEC) protocol. A) DNA with a known sequence of interest, B) A biotinylated primer is designed to anneal to the known sequence and linear amplification is performed. The biotinylated primer is extended into the adjacent DNA. C) The reaction is added to streptavidin beads to enrich for the biotinylated DNA. D) The biotinylated DNA is eluted from the streptavidin beads by heating.

The species’ present in a community may also be defined by amplifying part of a single locus, conserved in all bacteria and archaea that acts as a means of defining species, phylogenetic marker genes. In this type of analysis, genomic DNA is isolated from an environmental sample and is PCR amplified with primers specific for part of a phylogenetic marker gene. Base differences in these genes identify different phylogenetic groups. The amplified genes are sequenced to define the diversity of the community, but there are biases inherent to PCR that affect the outcome. PCR can be influenced by inhibitors in a sample, differential amplification, and chimera formation (V. Wintzingerode, Göbel et al. 1997, Haas, Gevers et al. 2011, Schloss, Gevers et al. 2011). Humic acids commonly found in soils inhibit Taq DNA polymerase at
concentrations as low as 0.08 μg/mL (Tebbe and Vahjen 1993). Differential amplification occurs when there are unequal annealing efficiencies between the primers used and the DNA template (Suzuki and Giovannoni 1996). These differences can limit amplification and result in failure to detect organisms present. Chimeras are artifacts that form when two different DNA molecules self-anneal during PCR and replicate creating a product that does not exist. Chimeras are especially problematic when PCR amplifying genes that share multiple conserved regions like the 16S rRNA gene, resulting in overestimates of the true microbial diversity (Wang and Wang 1997, Haas, Gevers et al. 2011).

Organism identification depends on successful amplification of phylogenetic gene markers. Conserved regions of phylogenetic marker genes, such as the 16S rRNA gene, are used to design primers. These sites are not well conserved in all taxa and can result in poor or unequal amplification (Suzuki and Giovannoni 1996, Jumpstart Consortium Human Microbiome Project Data Generation Working 2012). Primer pair selection can affect the amplification of templates due to the additive effect of biases in each primer to the template (Klindworth, Pruesse et al. 2013). In addition, large differences in guanine and cytosine (GC) content can also alter the success of PCR amplification for GC rich organisms (Aird, Ross et al. 2011). This has been observed during PCR amplification of genomic DNA from both AT and GC rich organisms in preparation for high-throughput sequencing on Illumina platforms (Aird, Ross et al. 2011).

To be useful as a phylogenetic marker gene, a locus must meet the following criteria: be present in all organisms, and be of sufficient length to be statistically significant with slowly and rapidly evolving sections to provide enough differences to distinguish one organism from another (Woese 1987). The 16S rRNA gene meets all these criteria. It is present in all prokaryotes, is roughly 1,500 bases long, and contains conserved and variable regions to track
evolutionary changes. There are a total of nine variable regions that range in size from 50-100 bases in length (Baker, Smith et al. 2003). The 16S rRNA gene encodes for a ribosomal RNA molecule that, along with 21 proteins, is part of the 30S small subunit of the prokaryotic ribosome (Mizushima and Nomura 1970). This gene was first suggested to be used as a phylogenetic marker gene in 1977 (Woese and Fox 1977) and is still the most commonly used marker today.

The specific protocols used for defining prokaryotic community composition through characterization of the 16S rRNA gene sequence have changed a great deal since 1977, but the overall approach remains the same. Genomic DNA is isolated and the 16S rRNA genes present are enriched in a manner that is compatible with the method used to sequence the DNA. Today, high-throughput sequencing (HTS) using the Illumina platform is emerging as the method of choice for sequencing due to its low cost and high sequencing output (Glenn 2011). The Illumina platform HiSeq 2000 can generate over 50 gigabases of sequencing data per day (Caporaso, Lauber et al. 2012) which enables users to multiplex several projects simultaneously while still generating large amounts of sequencing data for each project. Illumina platforms rely on sequence-by-synthesis technology and developed iTags for amplicon sequencing of the 16S rRNA gene (Degnan and Ochman 2012). iTags are PCR amplified portions of the 16S rRNA gene (Illumina®, San Diego, California). The portions can be either a single or multiple variable regions (~300 bases). The PCR amplified product then undergoes library preparation and sequencing (Illumina®, San Diego, California).

Different variable regions within the 16S rRNA gene can be sequenced with the region used ranging from 100-400 bases. Amplicon sequencing of the 16S rRNA gene is a convenient means of obtaining a snapshot of the prokaryotic diversity in addition to detecting members of
the “rare biosphere” – organisms present at low abundance that would otherwise be missed when sequencing bulk DNA (Sogin, Morrison et al. 2006). Protocols have been developed for amplicon sequencing of a single (Caporaso, Lauber et al. 2012) or multiple variable regions of the 16S rRNA gene (Jumpstart Consortium Human Microbiome Project Data Generation Working 2012). However, this approach does have some drawbacks. PCR amplification is a necessary step prior to library preparation and thus, the PCR biases previously mentioned might be introduced. In addition, primer pairs that amplify the 16S rRNA gene can introduce amplification biases both in the ability to amplify different organisms, i.e. some phyla may be missed (Klindworth, Pruesse et al. 2013) and abundance estimations (Jumpstart Consortium Human Microbiome Project Data Generation Working 2012).

**Objectives of This Study**

The goal of this study was to develop a method that would selectively capture specific DNA sequences from a complex mixture, providing higher efficiency and longer captured DNA sequences compared to previous biotin-streptavidin capture techniques and eliminate primer pair biases by using a single primer to capture specific loci. The method has been applied to capture the 16S rRNA genes from three microbial communities, an artificial bacterial mock community composed of twenty species (Chapter 3), three marine environmental samples: seawater, sediment, and the Easter oyster, *C. virginica* (Chapter 4), and a low biomass metagenome from basal ice of Taylor Glacier in Antarctica (Chapter 5). In addition, in Chapter 6, SEC is used to facilitate the characterization of single insertion mutants in an insertion library of the unicellular alga, *Chlamydomonas reinhardtii* by capturing DNA adjacent to insertions in an effort to determine their genomic locations.
CHAPTER 2.
METHODS

SEC Capture Primers Design

The SEC protocol requires a biotinylated primer and several primers were designed using the software Primer-BLAST (Ye, Coulouris et al. 2012). Table 2.1 lists each primer used in this work. This table includes biotinylated primers designed to capture the gltS and 16S rRNA gene (gltS capture and biotinylated Uni1390R) discussed in Chapters 3, 4, and 5, and the AphVIII cassette (Gonzalez-Ballester, de Montaigu et al. 2005) described in Chapter 6.

Linear Amplification Reactions

The capture primers initiate the linear amplification reaction. All linear amplification reactions used in this work were combined with AccuTaq™ DNA polymerase (0.05 U/µL, final concentration, Sigma-Aldrich, Saint Louis, MO), extreme thermostable single-stranded binding protein (ET SSB, 4 ng/µL, final concentration, New England Biolabs, Ipswich, MA), biotinylated capture primer (2 µM, final concentration), and deoxyribonucleotides (dNTPs, 200 µM, final concentration, Sigma-Aldrich, Saint Louis, MO) in a total volume of 50 µL. Each linear amplification reaction was placed in a BioRad C1000 thermal cycler (Bio-Rad Laboratories, Hercules, CA) programed for (a) 95°C for three minutes, (b) 95°C for forty-five seconds, (c) 60°C for thirty seconds plus an additional fifteen seconds added to every cycle, (d) 72°C for one and a half minutes, with a final extension step at 72°C for five minutes. Steps (b-d) were repeated for sixty times. Because a linear amplification is performed, sixty cycles were used to increase biotinylated product concentration.
Table 2.1 List of primers used in thesis.

<table>
<thead>
<tr>
<th>Primer name</th>
<th>Primer Forward</th>
<th>Primer Reverse</th>
<th>Chapter</th>
</tr>
</thead>
<tbody>
<tr>
<td>gltS capture</td>
<td>biotin-GTGATGC GGATACAAAGGAGT</td>
<td>biotin-GACGGGGCGGTGTTGTA CAA</td>
<td>3</td>
</tr>
<tr>
<td>biotinylated Uni1390R</td>
<td></td>
<td>biotin-GACGGGGCGGTGTTGTA CAA</td>
<td>3, 4, 5</td>
</tr>
<tr>
<td>AphVIII capture</td>
<td>biotin-GGCGGTGGATGGAAGATAC</td>
<td></td>
<td>6</td>
</tr>
<tr>
<td>291b gltS</td>
<td>ACGCTGACGTTGCTGCTCGG</td>
<td>ATGCCAACCACACGCCCCACC</td>
<td>3</td>
</tr>
<tr>
<td>1035b gltS</td>
<td>TGTGGGAGCTGGATGGAAGATAC</td>
<td>CCACAGTGACC CGCAG CCCAG</td>
<td>3</td>
</tr>
<tr>
<td>6207b gltS</td>
<td>ATGCAGCAGCGCAGCCATCA</td>
<td>CGTCTGCAGCAGCGCCTATCT</td>
<td>3</td>
</tr>
<tr>
<td>dapA</td>
<td>CAGAAGGGCATTTTGCCGAG</td>
<td>ATGCCCATTTCA CCGGATT</td>
<td>3</td>
</tr>
<tr>
<td>S-D-Bact-0008-d-S-20</td>
<td>AGAGTTTGATCCTGGGCTCATG</td>
<td></td>
<td>3, 4, 5</td>
</tr>
<tr>
<td>S-D-Bact-0337-a-S-20</td>
<td>CTCCTACGGGAGGCGAGCA</td>
<td></td>
<td>3</td>
</tr>
<tr>
<td>U926</td>
<td></td>
<td>CCGICIAATTIITTIAGTTT</td>
<td>3</td>
</tr>
<tr>
<td>AphVIII insertion</td>
<td>TCCTCCGTTGATTTTGGCCT</td>
<td>CTCCCCCCGTTCGCTGCTGAT</td>
<td>6</td>
</tr>
<tr>
<td>Cre09.g405750-A</td>
<td>TCCTCCGTTGATTTTGGCCT</td>
<td>CCGAGTCGCTTGGTTTTGCGA</td>
<td>6</td>
</tr>
<tr>
<td>Cre09.g405750-B</td>
<td>TCCTCCGTTGATTTTGGCCT</td>
<td>TCAGCGTGA TGACAGCA</td>
<td>6</td>
</tr>
<tr>
<td>Cre09.g410050</td>
<td>TCCTCCGTTGATTTTGGCCT</td>
<td>GGCTTTTTCCACACATGACCGG</td>
<td>6</td>
</tr>
<tr>
<td>cblp gene</td>
<td>TCGGAGTCCA ACTACGGCTA</td>
<td>CATCGGAGGAGATGACCACG</td>
<td>6</td>
</tr>
</tbody>
</table>

Note: Color designations can be found in Chapter 6.
Linear amplification reactions executed in Chapter 5 used an annealing temperature of 53°C instead of 60°C after a PCR temperature gradient was performed using the biotinylated Uni1390R and S-D-Bact-0008-d-S-20 (Table 2.1) revealed the temperature of 53°C gave the best results. Template DNA added for each linear amplification reaction varied based on experiment. Template DNA added for each linear amplification reaction varied from study to study.

Chapter 3: Linear amplification reactions involving the gltS capture primer (Table 2.1) used isolated E. coli genomic DNA (6 ng/µL, final concentration). Capture of the 16S rRNA gene using the biotinylated Uni1390R (Table 2.1) used isolated genomic DNA from the mock community HM-277D (1.1 ng/µL, final concentration). Chapter 4: Though triplicate DNA extractions of oyster, sediment, seawater and depurated oysters were performed, only a single DNA extraction from each sample was used to perform linear amplification reactions with biotinylated Uni1390R. Linear amplification reactions were performed in triplicate on a single DNA extraction from oyster (76 ng/µL, final concentration), sediment (16 ng/µL, final concentration), seawater (2 ng/µL, final concentration), and depurated oyster (82 ng/µL, final concentration) sample. Chapter 5: A total of five linear amplification reactions were performed using the biotinylated Uni1390R primer (Table 2.1) with isolated genomic DNA from the basal ice of Taylor Glacier at a final concentration of 2.1 pg/µL (105 pg per linear amplification reaction). Chapter 6: Linear amplification reactions were performed using isolated C. reinhardtii genomic DNA from individual strains or pooled library DNA. The amount of genomic DNA added to each SEC reaction of individual mutants CAH-8A, CAH-8B, and ATPase were 950 ng, 1,232 ng, and 780 ng, respectively. Linear amplification reactions were performed with the addition of genomic DNA at concentrations of 1.6 ng/µL for the 180 pool, and 10 ng/µL for the 1,440 pool in a 50 µL reaction.
Retrieving the Captured Fragments

Dynal M-280 streptavidin-coated magnetic beads (KilobaseBINDER™ kit, Life Technologies, Grand Island, NY) were used to capture the biotinylated DNA created by linear amplification. The protocol followed the manufacturer’s directions with some modifications. The entire 50 µL of the linear amplification reaction was added to the 150 µg of Dynal beads and slowly rotated (6 revolutions per minute) for three hours at 22°C. The Dynal bead-bound biotinylated DNA was washed five times with the washing solution provided by the manufacturer, followed by two rinses with deionized water (diH2O).

For retrieval of 16S rRNA genes (Chapters 3, 4, and 5), the biotinylated DNA was made double-stranded by combining the bead-bound DNA with 32.6 µL of diH2O supplemented with AccuTaq™ DNA polymerase (0.05 U/µL, final concentration), ET SSB (4 ng/µL, final concentration), S-D-Bact-0008-d-S-20 (Table 2.1, 2 µM, final concentration) and dNTPs (200 µM, final concentration) before extension. Extensions were performed in a BioRad C1000 thermal cycler programmed for (a) 95°C for three minutes, (b) 60°C for one minute, and (c) 72°C for ten minutes.

For retrieval of biotinylated DNA from a C. reinhardtii insertion mutant (Chapter 6), the biotinylated DNA was combined with 32.6 µL of diH2O supplemented with AccuTaq™ DNA polymerase (0.05 U/µL, final concentration), ET SSB (4 ng/µL, final concentration), random primers (60 ng/µL, final concentration, Promega, Madison, WI), and dNTPs (200 µM, final concentration) and were placed in a BioRad C1000 thermal cycler programmed for (a) 95°C for three minutes, (b) 60°C for one minute, and (c) 72°C for ten minutes.

Post extension, the Dynal beads/biotinylated DNA were transferred to a microcentrifuge tube where beads were subjected to three washes with washing solution and two washes with
diH₂O. To elute the double-stranded biotinylated DNA, 20 µL of diH₂O was added to the reaction-bead mixture and heated to 95°C for three minutes. A magnet (MagneSphere®, Promega, Madison, WI) was used to remove the beads; the supernatant contains double-stranded biotinylated DNA.

**The Mock Community**

Purified mock community (HM-277D) genomic DNA (56 ng/µL) was obtained from BEI Resources (through the American Type Culture Collection, Bethesda, MD) and used in Chapter 3. This mixture of genomic DNAs includes contributions from 20 bacterial species in varied abundances as will be seen in Chapter 3 in a table and figure. The mock community DNA was stored at -80°C.

**Quantitative PCR (QPCR) Amplifications**

All QPCR reactions were performed with SYBR® Select Master Mix (Life Technologies, Grand Island, NY) was combined with sample DNA and the forward and reverse primer (0.4 µM, final concentration) for each site. An ABI 7000 PCR system (Life Technologies, Grand Island, NY) was used for each reaction with the following conditions: (a) 95°C for three minutes, (b) 95°C for ten seconds, (c) 60°C for thirty seconds. Steps (b-d) were repeated for forty times. QPCR of *gltS* gene (Chapter 3): QPCR primers were designed to amplify fragments beginning 63, 1,035, and 6,207 bases downstream of the *gltS* capture primer-binding site (Table 2.1), producing PCR amplicons of the sizes 247, 151, and 101, respectively. An additional primer set amplified the *dapA* gene to measure for non-specific carryover of genomic DNA. Standard curves were generated for each primer set using purified *E. coli* MG1655 genomic DNA at known concentrations from 3-300 ng/µL. QPCR on *C. reinhardtii* loci (Chapter 6): Two primer sets were designed within *C. reinhardtii*, one to amplify a single insertion in the gene
Cre09.g410050 and another to amplify using the cblp gene (g6364) (Table 2.1). Both primer sets were designed using Primer- BLAST (Ye, Coulouris et al. 2012) and produced PCR products less than 250bp in length listed in Table 2.1. Standard curves were generated using purified wild type C. reinhardtii genomic DNA.

**PCR Amplifications**

**PCR of 16S rRNA Gene V3-V5 Variable Regions (Chapter 3):**

The PCR was performed on genomic DNA from the HM-277D mock community that amplified the V3-V5 regions of the 16S rRNA gene. Genomic DNA from the MC (0.8 ng/μL, final concentration), AccuTaq™ DNA polymerase (0.05 U/μL, final concentration), S-D-Bact-0337-a-S-20 (0.4 μM, final concentration), U926 (0.4 μM, final concentration), ET SSB (4 ng/μL, final concentration), and dNTPs (200 μM, final concentration) were combined in 50 μL reactions. PCR was performed using a Bio-Rad C1000 thermal cycler programmed as follows: (a) 95°C for three minutes, (b) 95°C for forty-five seconds, (c) 60 °C for thirty seconds, (d) 72°C for forty-five seconds, and a final extension step (3) 72°C for five minutes. Steps (b-d) were repeated thirty-five times. The PCR product was cleaned using Gene Clean® Spin Kit (MP Biomedicals, Santa Ana, CA). The purified product was sequenced using an Illumina HiSeq 2000 platform.

**PCR on SEC Captured DNA from Individual C. reinhardtii Mutants (Chapter 6):**

A single primer was designed to anneal 56 bases upstream of the 3´ end of the insertion for each allele (Table 2.1 and Figure 2.1). Primers were also designed to specifically to amplify downstream of the insertions in Cre09.g405750 and Cre09.g410050 (Table 2.1, Figure 2.2). Post SEC, these primers were used to confirm SEC captured DNA adjacent to each insertion. The biotinylated DNA was combined with a hot start AccuTaq™ DNA polymerase (0.05 U/μL, final
concentration), insertion forward primer (400 nM, final concentration), mutant primer (400 nM, final concentration), and dNTPs (200 µM, final concentration) in a total volume of 50 µL. A BioRad C1000 thermal cycler was used and programed for: (a) 95°C for three minutes, (b) 95°C for forty-five seconds, (c) 60º for thirty seconds, (d) 72°C for forty-five seconds, and a final extension step (3) 72°C for five minutes. The cycle (b-d) was repeated for forty times.

Figure 2.1 A) AphVIII insertion used for generation of insertion library. B) Primer locations for SEC, PCR, and quantitative PCR. Black line indicates location of the biotinylated capture primer located 180 bases from the end of the RBCS2 terminator sequence. Purple arrow indicates primer designed 56 bases from end of AphVIII gene insertion called the insertion forward primer. Color designations for primers are same as found in Table 2.1
Figure 2.2 Primer locations for three individual insertion mutants used for detection of SEC capture using PCR and quantitative PCR. A) Primer locations for CAH8 insertions mutants. Blue arrow indicates location of the reverse primer for CAH8-A insertion mutant and green arrow the reverse primer for CAH8-B insertion mutant. B) Reverse primer location for ATPase insertion mutant located by orange arrow. Color designations for primers are same as found in Table 2.1

**Illumina HiSeq 2000 Sequencing**

Chapter 3: Biotinylated DNA obtained from SEC capture of the mock community (SEC-MC) and PCR amplification of V3-V5 regions of the 16S rRNA genes from the mock community (PCR-MC) were sequenced using 1/3 of a lane each on an Illumina HiSeq 2000 with 100 cycles, paired-end sequencing. The initial DNA concentration for the SEC-MC and PCR-MC was 0.41 ng/µL and 8.9 ng/µL, respectively. The University of Minnesota Genomics Center was contracted to create the library preparations using a TruSeq ChIP DNA kit (Illumina, Inc., San Diego, CA). Libraries were fragmented to ~270 base pairs and post-fragmentation size selection concentrations were 8.1 ng/µL and 63.5 ng/µL, respectively as determined by Agilent High Sensitivity DNA Kit (Agilent Technologies, Santa Clara, CA). Chapter 4: Post SEC, the triplicate SEC reactions from each sample (oyster, sediment, seawater, and depurated oyster) were pooled together for library preparation and DNA sequencing. These were termed SEC-
Oyster, SEC-Sediment, SEC-Seawater, and SEC-DepuratedOyster. DNA concentrations for each were as follows: SEC-Oyster 0.3 ng/µL, SEC-Sediment 0.2 ng/µL, SEC-Seawater 0.1 ng/µL, and SEC-DepuratedOyster 0.2 ng/µL. The University of Georgia performed library preparation using Nextera XT DNA Library Prep kit. Libraries were fragmented to ~250 base pairs and size selection was performed with a DN-486 High Sensitivity NGS Fragment Analysis Kit. DNA library concentrations post-size selections were as follows: SEC-Oyster 2.9 ng/µL, SEC-Sediment 4.3 ng/µL, SEC-Seawater 4.3 ng/µL, and SEC-DepuratedOyster 3.9 ng/µL. Libraries were paired-end sequenced for 100 cycles and sequenced on a single lane on an Illumina HiSeq 2000. Chapter 5: All five SEC reactions were combined together prior to library construction to increase the DNA concentration, now designated SEC-TG. The initial DNA concentration for the SEC-TG was 0.1 ng/µL. Library preparation and DNA sequencing was performed by the University of Wisconsin-Madison Biotechnology Center using Illumina’s ChIP Sample Prep kit (Illumina, Inc., San Diego, CA) and Illumina HiSeq 2000 v1.5. Libraries were fragmented to ~150 base pairs using a Bioruptor® (12 minutes-15 seconds on, 15 seconds off). Post-fragmentation, size selection was performed by agarose gel electrophoresis. DNA library concentration post-size selection was 2.3 ng/µL.

**Trimming of Illumina Datasets**

All Illumina datasets were trimmed using the bioinformatic software CLC Genomics Workbench version 6.0.4. The following trimming parameters were applied: quality score limit of 0.05, and no ambiguous base. Sequence read lengths below 95 (Chapters 4 and 6), 90 (Chapter 3), and 45 bases (Chapter 5) were discarded.
Read Mapping of Datasets

In Chapter 3, SEC-MC and PCR-MC datasets were mapped to the 16S rRNA gene sequences of each bacterial strain from the HM-277D mock community using CLC Genomics Workbench version 6.0.4. To obtain sequence reads that were specific to each species, the following parameters were used: length fraction of 0.97, similarity fraction of 0.97, mismatch cost of 2, insertion cost of 3, deletion cost of 3, non-specific matches were ignored. Under these parameters, mapped sequence reads were considered species-specific. Additional relaxed read mappings were done with the same parameters with the exceptions of length of fraction and similarity fraction set at 0.80. Relaxed read mapping parameters were performed to account for sequence reads that were not species-specific.

Generation of Simulated Datasets

In Chapter 3, the program EMIRGE was tested with datasets that contained sequence reads mapping to only 16S rRNA gene sequences like those generated following SEC. Simulated datasets were generated composed of random 100 bases, error-free sequence reads of reference 16S rRNA genes from the 20 organisms found in the HM-277D using the program wgsim (Li, Handsaker et al. 2009). Only one 16S rRNA gene from each organism was used to generate the simulated datasets regardless of 16S rRNA gene copy number in their respective genomes. Multiple simulated datasets were generated that varied in total number of sequence reads and number of sequence reads per 16S rRNA gene from organisms.

QIIME and EMIRGE Analysis

The QIIME (Quantitative Insights Into Microbial Ecology) pipeline combines multiple programs together for analyzing large datasets specific to microbial ecology (Caporaso, Kuczynski et al. 2010) and was used to V4 iTags in Chapter 5.
All QIIME analysis provided in this thesis was completed using the de novo OTU picking method (*pick_de_novo_otus.py*) based on the accuracy when tested with simulated data. The workflow for de novo OTU picking is broken down into four stages: (i) sequence reads are clustered into operational taxonomic units (OTUs) using the program UCLUST. OTUs are a way to assess microbial diversity, by clustering 16S rRNA gene sequences that share a given percentage similarity. (ii) A representative sequence from each OTU cluster is chosen based on the centroid sequence, the sequence present in the highest abundance in the cluster, (iii) the representative sequences are then aligned using the tool PyNAST (Caporaso, Bittinger et al. 2010), (iv) taxonomy is assigned to representative sequences using UCLUST, (v) alignment files are filtered and (vi) a phylogenetic tree is built using the program FastTree (Price, Dehal et al. 2010). An additional step of using the Ribosomal Database Project to assign taxonomy to representative OTUs using the naïve Bayesian classifier is also performed (Wang, Garrity et al. 2007). The output of QIIME’s de novo OTU picking workflow results in the identification of OTUs for microbial diversity inferences. QIIME workflows were designed for high-throughput amplicon sequencing of the 16S rRNA gene. QIIME analyses were performed on a high performance-computing cluster on a single node with four processors.

Another program used in this thesis, Expectation Maximization Iterative Reconstruction of Genes from the Environment (EMIRGE) is designed to reconstruct full-length 16S rRNA gene sequences from high-throughput datasets (Miller, Baker et al. 2011). EMIRGE uses the expectation maximization algorithm and through iterative cycles, attempts to recreate full-length 16S rRNA gene sequences from short sequence reads. For each iteration, EMIRGE performs the following functions: (i) sequence reads are mapped to a reference database, (ii) the probability that the reference sequence generated a sequence read is calculated, (iii) the abundance of that
reference is calculated, and (iv) the reference sequence is corrected by the bases in reads that have the highest probability (Miller, Baker et al. 2011). These steps are repeated over 40 iterations and gradually the reference sequences are corrected with the sequence reads and their abundances are determined by the number of sequence reads attributing to the assembled sequence. EMIRGE requires input datasets to be in fastq format to account for any sequencing errors in the dataset.

Due to computational limitations, all datasets were subsampled with a custom script (fastq_random_subset.py) (Miller 2013) for analysis using EMIRGE. All datasets analyzed with EMIRGE used the emirge_amplicon.py script with default parameters and a maximum read length of 101 bases. All analyses (Chapters 3, 4, and 5) were performed on a high performance-computing cluster with a large memory capacity of 1 TB on a single node with either 10 or 20 processors. The assembled sequences generated by EMIRGE were used to query the Silva database (SSU_NR_111) using BLAST v.2.2.28. The top hits from BLAST based on bit-scores were chosen to distinguish the taxonomic classification of each assembled sequence output from EMIRGE.

Both programs are designed to analyze HTS datasets to provide taxonomic classification of bacteria whether from PCR amplicon 16S rRNA sequencing datasets (QIIME), or metagenomic datasets (EMIRGE) to determine bacterial diversity.

Statistical Analyses

Statistical analyses in this thesis were performed using the StatPlus®:mac statistical analysis software (AnalystSoft Inc, Alexandria, VA). In Chapter 3, F-test was used to test for equal variances for the mean copy number of retrieved molecules at Sites 1, 2, and 3 to the mean copy number of *dapA* gene controls. Student’s t-test was used compare the mean copy number
of retrieved molecules at Sites 1, 2, and 3 to the mean of the dapA gene controls. If variances were unequal, a heteroscedastic Student’s t-test was performed.

**Oyster, Sediment, and Seawater Sample Collection**

In Chapter 4, oyster, sediment, and seawater samples were collected from the Louisiana Sea Grant Oyster Hatchery in Grand Isle, Louisiana on July 15, 2013. An overview of the experimental scheme is provided in Figure 2.3. Sixteen oysters were collected from a single oyster cage located approximately 50 feet from shore. Oysters were placed in burlap bags and kept out of sunlight during transport to Louisiana State University for processing. Approximately 7 hours later, the outside of each oyster shell was scrubbed using wire brushes to remove excess sediment and rinsed with 70% ethanol. Eight oysters were placed in a depuration tank with salinity at 20 parts per thousand (ppt) for seven days. Sediment samples were collected in triplicate next to the oyster cage using a dredge (Wildlife Supply Company model number 146-B12-1096). Approximately 50 grams of sediment was transferred in separate sterile conical tubes in triplicate. The samples were stored at room temperature out of sunlight and transported back to Louisiana State University for DNA extractions. Seawater was collected (approximately 800 mL) within six inches of the oyster cage by submersion of an autoclaved 1-liter bottle then opening bottle to fill approximately 800 mL at a depth of 2 feet from the surface. Bottles were then wrapped, kept at room temperature, and transported back to Louisiana State University for filtration and DNA extractions. Salinity and water temperature was measured using a YSI Model 85 less than one foot away from the oyster cages and was 20.2 ppt and 30.1 °C.
DNA Extraction of Oysters, Sediment, and Seawater

In Chapter 4, seven hours post–harvest, three oysters were chosen from the burlap bag and washed with 70% ethanol to remove any contaminating organisms present on the outside of the shells. The oysters were then shucked using a sterile shucking knife. Each oyster was left intact and placed in a sterile 50 mL conical tube and weighed. The average mass of the oysters was 10.6 ± 2.7 grams. Individual oysters were homogenized separately using a handheld tissue homogenizer with equal volumes of PBS (w/v). DNA extractions on each homogenized oyster was performed in triplicate using PowerSoil® DNA Isolation Kit (MO BIO, Carlsbad, CA) with the addition of 250 μL of whole oyster homogenate. DNA was extracted from each triplicate sediment sample by combining five grams of sediment with equal volumes of PBS (w/v) and agitated using a vortex mixer for five minutes. DNA extractions were performed in triplicate for
each sediment sample by adding 200 μL of mixture to PowerSoil® DNA Isolation Kit. Twenty-four hours post-collection, 100 mL of collected seawater was filtered using a 0.22 μm polycarbonate filter (diameter 47 mm, Sterlitech, catalog number PCTF0247100) in triplicate. Each filter was cut in half with one half being placed in PBS and agitated using a vortex mixer to remove cells from the filter surface. The remaining half of the filter was further cut into thirds and underwent DNA extraction. Each filter was cut in half using a sterile scalpel. Half of filter was cut into thirds and DNA was extracted from approximately 16.6 mL of seawater using the PowerSoil® DNA Isolation Kit.

Eight oysters were placed in a depuration tank. This tank held 400 liters of artificial seawater that circulated through a 10 μm and 2 μm and was then UV sterilized. The depuration tank water had a salinity of 20 ppt. After seven days in this tank, three depurated oysters were chosen and rinsed with 70% ethanol. Depurated oysters were shucked and weighed with an average mass of 10.4 ± 4.0 grams. Three depurated oysters were separately homogenized in equal volumes of PBS (w/v). DNA extractions for each homogenized depurated oyster was performed in triplicate by adding 250 μL of whole oyster homogenate to the PowerSoil® DNA Isolation Kit.

The C. reinhardtii Insertion Library

The C. reinhardtii insertion library was generously provided by the laboratory of James V. Moroney (Department of Biological Sciences, Louisiana State University) in Chapter 6. This library contains 32,000 single colony isolates that were isolated in the presence of paromomycin on tris-acetate-phosphate (TAP) plates; each plate containing 180 isolates. The Moroney laboratory performed all genomic DNA extractions from the insertion library. Genomic DNA extracted from a single TAP plate is referred to as a “180 pool”. Genomic DNA extracted from
eight individual TAP plates is termed a “1,440 pool”, and contained genomic DNA from 1,440 single colony isolates.

**Characterized *C. reinhardtii* Insertion Mutants**

The Moroney laboratory also provided genomic DNA from three well-characterized insertion mutants for Chapter 6. Two of the three individual mutants contained a single insertion into the carbonic anhydrase 8 gene (CAH8, Cre09.g405750); in strain CAH8-A, the insertion is located in the first intron of CAH8 gene, in strain CAH8-B, the insertion is located in the 5´ untranslated region (UTR) of CAH8. The third mutant carries an insertion located in the last exon of Cre09.g410050, a Ca^{+2}-transporting ATPase gene.
CHAPTER 3.
SELECTED ENRICHMENT THROUGH CAPTURE (SEC): METHOD FOR RETREIVING SINGLE LOCI FROM COMPLEX MIXTURES

Introduction

Selectively sequencing the 16S rRNA genes isolated from bacterial metagenomes is an experimental approach routinely implemented when defining diversity in environmental samples. The success of this approach, as defined by the ability to thoroughly characterize a bacterial community, has improved with advancements in high-throughput sequencing technologies compared to Sanger sequencing (Shendure and Ji 2008). An experimental paradigm currently employed utilizes a set of primers that anneal to conserved regions of the 16S rRNA gene that are used to amplify variable regions within the loci (Sogin, Morrison et al. 2006, Caporaso, Lauber et al. 2012, Lie, Liu et al. 2014). The resulting fragments are directly sequenced (Sogin, Morrison et al. 2006, Degnan and Ochman 2012), and various analyses are conducted to reconstruct community phylogeny. Assuming that individual 16S rRNA genes are amplified with equivalent efficiencies, direct sequencing at sufficient depth offers the promise of defining the individual members of the community. Unfortunately, these approaches are potentially biased because they rely on PCR amplification and as discussed in Chapter 1, there are a number of PCR-based artifacts that can confuse the interpretation of microbial community structures built around such data (Jumpstart Consortium Human Microbiome Project Data Generation Working 2012, Lee, Herbold et al. 2012, He, Zhou et al. 2013).

In addition to PCR biases, the length of amplified 16S rRNA gene can impact both the species richness and evenness (Youssef, Sheik et al. 2009, Jumpstart Consortium Human Microbiome Project Data Generation Working 2012, Yarza, Yilmaz et al. 2014). Individual variable regions of the 16S rRNA gene, such as V1, V2, and V6 can overestimate the number of
species present in a given sample while V4 and V7 underestimate the number of species (Youssef, Sheik et al. 2009). Analysis of species level diversity of each individual variable regions V2-V6 were shown to underestimate species richness when compared to full-length 16S rRNA gene sequences, while V1 overestimated the number of species (Yarza, Yilmaz et al. 2014). Multiple variable regions also negatively impact the accurate descriptions of the microbial community. Variable regions V3-V5 were shown to be more accurate in defining a simple mock community when compared to V1-V3 and V6-V9, however V3-V5 still under- and overestimated species evenness when compared to full-length 16S rRNA gene sequences (Jumpstart Consortium Human Microbiome Project Data Generation Working 2012).

With this work, an attempt is made to develop a method that limits PCR bias and retrieves full-length 16S rRNA genes to provide a better description of a microbial community versus PCR amplification of V3-V5. The method termed Selective Enrichment through Capture (SEC) utilizes biotinylated primers to retrieve specific DNA sequences from mixtures of DNA. Once isolated, this sequence is repeatedly copied using a single primer and a thermostable DNA polymerase, amplifying the captured sequence by creating a large number of single-stranded DNA fragments. While the technique requires primer binding and extension by a polymerase, it is not a polymerase chain reaction as the two primers are not used simultaneously. Replicating the sequence in this manner should limit the impact of some PCR biases, particularly amplification biases caused by primer pair-template mismatches and abundance estimation that can affect assessing community composition (Jumpstart Consortium Human Microbiome Project Data Generation Working 2012, Klindworth, Pruesse et al. 2013). The technique provides additional copies of the targeted sequence through linear amplification for downstream applications.
This chapter documents the efficiency of SEC in retrieving a targeted locus from within a single genome, demonstrating the technique is capable of recovering the entire gene of interest, and as much as 5,000 bases of adjacent sequence as well. In addition, SEC is successfully used to target the 16S rRNA gene of 20 bacteria found in an artificially generated mock community. The capture primer was designed to anneal to an 18 base pair sequence conserved within this group of bacteria. Recovered DNA fragments were subjected to high-throughput sequencing and analysis of the resulting dataset identified all twenty bacteria. In addition, relative abundances of this mock community were recreated within an order of magnitude establishing the potential utility of SEC in characterizing bacterial community structure.

Results

SEC Retrieval of a Single Locus

Initially an attempt was made to use SEC to recover a single locus from within that organism’s genome. For these studies, the gltS gene of *E. coli* MG1655 was targeted. This 1,206 base pair (bp) gene (map position 3,825,483 ↔ 3,826,688) encodes a sodium-dependent glutamate transporter (Kalman, Gentry et al. 1991). The gltS gene represents approximately 1/3700th of the genome, and is present at one copy per genome making it straightforward to quantify capture of this gene. Purified genomic DNA was mixed with a capture primer (Table 2.1) that annealed 25 bases upstream of the start site of the gltS gene. Quantitative PCR (QPCR) was used to monitor the effectiveness of the SEC protocol. Primers were designed to amplify fragments beginning 63, 1,035, and 6,207 bases downstream of the gltS capture primer-binding site and were designated as Site 1, 2, and 3, respectively (Table 3.1). These primers produced PCR amplicons of 247, 151, and 101 base pairs that were quantified. The gene was first targeted in 300 ng of genomic DNA. Assuming the DNA was composed of full-length genome
molecules (4.46 x 10^6 bp long), 300 ng corresponds to 6 x 10^7 genome copies available for capture. Table 3.1 reports the results of three independent SEC trials. Recovery was estimated at 2 x 10^9 gltS copies when fragments generated from primers amplifying Site 1 were evaluated, approximating the theoretical maximum of 3.6 x 10^9 copies available. (The maximum value was calculated by assuming that all 6 x 10^7 genome copies were linearly amplified for 60 cycles during the SEC protocol – each gltS copy increasing to two after a round of linear amplification.)

1 x 10^7 copies were recovered from Site 2, indicating an approximate 100-fold reduced efficiency of recovery with the 970 base pair increase in size of the captured fragment. SEC products were PCR amplified and subjected to Sanger sequencing. Results confirmed the amplified fragment sequences were identical to those reported for the corresponding regions of the MG1655 gltS coding sequence.

<table>
<thead>
<tr>
<th>Copy Number (x 10^7)</th>
<th>Site 1</th>
<th>Site 2</th>
<th>Site 3</th>
<th>dapA</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>210 ± 4.1</td>
<td>2.0 ± 0.22</td>
<td>0.27 ± 0.012</td>
<td>0.0083 ± 0.008</td>
</tr>
<tr>
<td>2</td>
<td>170 ± 3.6</td>
<td>0.57 ± 0.04</td>
<td>0.26 ± 0.006</td>
<td>0.048 ± 0.043</td>
</tr>
<tr>
<td>3</td>
<td>206 ± 11</td>
<td>0.25 ± 0.025</td>
<td>0.33 ± 0.02</td>
<td>0.0012 ± 0.0009</td>
</tr>
<tr>
<td>Mean</td>
<td>200 ± 20</td>
<td>0.95 ± 0.83</td>
<td>0.28 ± 0.035</td>
<td>0.019 ± 0.030</td>
</tr>
</tbody>
</table>

Note: Values in rows 1-3 are the mean copy number ± standard deviation of triplicate measurements. Values in the row labeled “Mean” are the averages of the copy numbers calculated for each site during the three independent trials performed in triplicate ± standard deviation.

Since the length of the captured fragment is only limited by the length of the template and the processivity of the DNA polymerase – the speed at which nucleotides are incorporated on the newly synthesized strand during DNA synthesis (Kelman, Hurwitz et al. 1998) – a third primer set was designed to detect and enumerate recovered sequences that included a site 5,000 bases.
downstream of the *gltS* termination codon (Site 3). This site was detected at levels approximately five-fold lower than what was observed at Site 2. This result suggests that SEC can be used to recover at least 6,200 bases of sequence adjacent to a capture primer, but indicates lower efficiencies as product length increases.

There is the possibility that a non-specific transfer of genomic DNA during SEC explains the results provided in Table 3.1. In sufficient quantity, genomic DNA would serve as a template for the QPCR reaction used to detect sequences recovered following SEC. To rule out this circumstance, primers specific for amplification of a sequence within the *dapA* gene of MG1655 (map position: 2,596,904 ↔ 2,597,782) were used to estimate carryover of genomic DNA during the SEC protocol (Table 3.1). The *dapA* locus was chosen because it is located 1.2 million bases from the *gltS* gene and should not be captured during SEC targeting of *gltS*. On average, approximately 2% of the total number of molecules recovered at Sites 1, 2, and 3 could be identified as the *dapA* gene (Table 3.1). The numbers of *gltS*-specific sequences detected were significantly greater (*P* < 0.05, Student’s t-test) than the *dapA* gene controls from all three sites (Site 1, 2, and 3), eliminating concern of non-specific carryover of genomic DNA affecting the outcome of the analysis.

Table 3.2 describes the effect of reduced concentrations of genomic DNA on fragment recovery during the SEC protocol. *E. coli* MG1655 genomic DNA was serially diluted to reduce the *gltS* copy number in the sample reaction. Following SEC, QPCR was performed with primers that amplify 1,035 bases from the capture primer-binding site (Site 2) to estimate the quantities of *gltS*-specific fragments recovered. When 30 ng (6 *×* 10⁶ genome copies) of genomic DNA is used, the protocol is as effective as was observed when using 300 ng DNA; the estimated numbers of recovered target DNA were near the theoretical maximum possible and
were significantly greater \((P < 0.05, \text{Student’s t-test})\) than the number of copies of \(\text{dap}A\) gene detected for all three trials. When copy number was lowered to \(6 \times 10^5\) genomes (equivalent to 3 ng), the reduction of starting material resulted in lower recoveries and greater variability in the amounts of biotinylated product recovered, as evidenced by the lower recovery in trial 2 relative to trails 1 and 3. However, the number of recovered \(\text{glt}S\)-specific fragments was significantly greater \((P < 0.05, \text{student’s t-test})\) than the \(\text{dap}A\) gene in all three trials. When the number of genome copies used in the reaction was reduced another ten-fold further \((300\ \text{pg}, 6 \times 10^4\ \text{genome copies})\), SEC was unable to reproducibly capture fragments in quantities that clearly distinguish the values obtained from possible carryover.

Table 3.2 The effect of reduced DNA concentrations of the effectiveness of SEC, as calculated from QPCR.

<table>
<thead>
<tr>
<th></th>
<th>30 ng</th>
<th>3 ng</th>
<th>0.3 ng</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Site 2 (\text{dap}A)</td>
<td>Site 2 (\text{dap}A)</td>
<td>Site 2 (\text{dap}A)</td>
</tr>
<tr>
<td>1</td>
<td>1.0 ± 0.06</td>
<td>0.003 ± 0.002</td>
<td>0.2 ± 0.014</td>
</tr>
<tr>
<td>2</td>
<td>3.9 ± 0.41</td>
<td>0.003 ± 0.003</td>
<td>0.03 ± 0.004</td>
</tr>
<tr>
<td>3</td>
<td>6.1 ± 0.35</td>
<td>0.003 ± 0.003</td>
<td>0.22 ± 0.017</td>
</tr>
<tr>
<td>Mean</td>
<td>3.7 ± 2.2</td>
<td>0.003 ± 0.003</td>
<td>0.15 ± 0.09</td>
</tr>
</tbody>
</table>

Note: Values in rows 1-3 are the mean copy number ± standard deviation of triplicate measurements. Values in the row labeled “Mean” are the averages of the copy numbers calculated for each site during the three independent trials performed in triplicate ± standard deviation.

The observations summarized in Tables 3.1 and 3.2 indicate that SEC facilitates the recovery of millions of copies of the targeted biotinylated product, and that fragments as large as 6,200 bp may be recovered with high efficiency. Clearly, if one is tasked with recovering a
single locus and/or sequence downstream of that locus, SEC offers an effective means of accomplishing this task.

**Combining SEC and High-Throughput Sequencing**

The success of retrieving a single locus with SEC led to the question: can multiple distinct but related sequences be captured from a mixture of genomic DNAs using the same biotinylated primer? To answer this question, an attempt was made to use SEC to retrieve 16S rRNA genes from a mock community.

The mock community, HM-277D (BEI Resources, ATCC, Bethesda, MD), is composed of 20 bacterial species. The relative abundance of genomic DNA from individual species within this mock community varies by as much as four orders of magnitude (Figure 3.1). A capture primer targeting near the 3´ end of the 16S rRNA gene was used to initiate SEC protocol. That primer replicates the sequence S-*:Univ-1390-a-A-18 (Uni1390R), an 18 base reverse primer targeting position 1407 → 1390 of the 16S rRNA gene of *E. coli* (Klindworth et al. 2013), differing only in that the 5´ end of Uni1390R is labeled with a biotin molecule (Table 2.1). The biotinylated Uni1390R perfectly complements the targeted region in 19 of the 20 bacteria included in HM-277D. There is a one base mismatch 12 bases from the 3´ end of this primer within the targeted region of the *Helicobacter pylori* 16S rRNA gene (Appendix Table 1). Following capture, the recovered fragments were made double-stranded using the primer S-D-Bact-008-dS-20 (Table 2.1) that is specific for a conserved sequence at the 5´ end of the 16S rRNA gene of the bacteria found in HM-277D.

The recovered fragments were directly sequenced using the Illumina HiSeq 2000 platform. Sequencing generated 110,986,206 raw paired-end reads, 101 bases in length. Raw sequence data were trimmed using the bioinformatic software package CLC Genomics
Workbench (CLC bio, Boston, MA), generating high quality score sequence reads for mapping. Post-trimming, this dataset consisted of 102,252,558 reads; this trimmed dataset is referred to here as SEC-MC, reflecting the use of SEC to characterize the mock community.

Species-specific reads from SEC-MC were mapped to the individual 16S rRNA genes using CLC Genomics Workbench (CLC) (Table 3.3). Since the number of single nucleotide polymorphisms between redundant 16S rRNA genes of species in HM-277D were not greater than 1% of the total number of bases in the gene, the sequence of a single 16S rRNA gene from each bacterium was used to represent that organism during mapping. Sequence reads were mapped to the individual 16S rRNA genes.

---

Figure 3.1 Reported and SEC-MC derived abundances of organisms within HM-277D. Reported abundances for organisms in HM-277D ranged from 0.02-26% and are shown in red. Several organisms contained the same relative abundances and are shown as a single dot. Abundance as determined by species-specific sequence reads from SEC-MC were plotted against the reported abundance (black).
Table 3.3 Abundances based on read counts from SEC captured 16S rRNA genes and PCR amplification of V3-V5 of the 16S rRNA genes from the HM-277D mock community.

<table>
<thead>
<tr>
<th></th>
<th>HM-277D</th>
<th>SEC-MC</th>
<th>PCR-MC</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Reported Abundance</td>
<td>Read Count</td>
<td>Calculated Proportional Abundance</td>
</tr>
<tr>
<td><em>Rhodobacter sphaeroides</em> ATCC 55304</td>
<td>26.0</td>
<td>1,018,256</td>
<td>3.0</td>
</tr>
<tr>
<td><em>Escherichia coli</em> MG1655</td>
<td>23.0</td>
<td>27,799,733</td>
<td>75</td>
</tr>
<tr>
<td><em>Staphylococcus epidermidis</em> ATCC 12228</td>
<td>20.0</td>
<td>2,443,150</td>
<td>7.0</td>
</tr>
<tr>
<td><em>Streptococcus mutans</em> ATCC 700610</td>
<td>20.0</td>
<td>3,114,300</td>
<td>8.0</td>
</tr>
<tr>
<td><em>Streptococcus agalactiae</em> ATCC BAA-611</td>
<td>2.0</td>
<td>639,045</td>
<td>1.7</td>
</tr>
</tbody>
</table>


(Table 3.3 continued)

<table>
<thead>
<tr>
<th></th>
<th>HM-277D</th>
<th>SEC-MC</th>
<th>PCR-MC</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Reported</td>
<td>Read</td>
<td>Calculated Proportional</td>
</tr>
<tr>
<td>Abundance</td>
<td>Abundance</td>
<td>Count</td>
<td>Proportional Abundance</td>
</tr>
<tr>
<td>Bacillus cereus ATCC 10987</td>
<td>2.0</td>
<td>705,998</td>
<td>1.9</td>
</tr>
<tr>
<td>Clostridium beijerinckii ATCC 51743</td>
<td>2.0</td>
<td>146,201</td>
<td>0.40</td>
</tr>
<tr>
<td>Pseudomonas aeruginosa ATCC 47085</td>
<td>2.0</td>
<td>160,223</td>
<td>0.40</td>
</tr>
<tr>
<td>Staphylococcus aureus ATCC BAA-1717</td>
<td>2.0</td>
<td>508,864</td>
<td>1.0</td>
</tr>
<tr>
<td>Helicobacter pylori ATCC 700392</td>
<td>0.20</td>
<td>69,727</td>
<td>0.20</td>
</tr>
<tr>
<td>Propionibacterium acnes KPA 171202</td>
<td>0.20</td>
<td>35,954</td>
<td>0.10</td>
</tr>
<tr>
<td>Acinetobacter baumannii ATCC 17978</td>
<td>0.20</td>
<td>138,276</td>
<td>0.30</td>
</tr>
</tbody>
</table>
(Table 3.3 continued)

<table>
<thead>
<tr>
<th></th>
<th>HM-277D</th>
<th>SEC-MC</th>
<th>PCR-MC</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Reported Abundance</td>
<td>Read Count</td>
<td>Calculated Proportional Abundance</td>
</tr>
<tr>
<td>Neisseria meningitides ATCC 13091</td>
<td>0.20</td>
<td>3,979</td>
<td>0.01</td>
</tr>
<tr>
<td>Listeria monocytogenes ATCC BAA-679</td>
<td>0.20</td>
<td>131,354</td>
<td>0.40</td>
</tr>
<tr>
<td>Lactobacillus gasseri ATCC 33233</td>
<td>0.20</td>
<td>8,561</td>
<td>0.02</td>
</tr>
<tr>
<td>Deinococcus radiodurans R1 ATCC 13939</td>
<td>0.02</td>
<td>3,312</td>
<td>0.01</td>
</tr>
<tr>
<td>Streptococcus pneumoniae ATCC BAA-334</td>
<td>0.02</td>
<td>15,590</td>
<td>0.04</td>
</tr>
<tr>
<td>Bacteroides vulgatus ATCC 8482</td>
<td>0.02</td>
<td>239</td>
<td>0.006</td>
</tr>
<tr>
<td>Enterococcus faecalis ATCC 47077</td>
<td>0.02</td>
<td>19,388</td>
<td>0.05</td>
</tr>
<tr>
<td>Actinomyces odontolyticus ATCC 17982</td>
<td>0.02</td>
<td>146,074</td>
<td>0.40</td>
</tr>
</tbody>
</table>

Note: Abundances values are reported as percent. Relative abundance of HM-277D mock community are reported according to number of 16S rRNA genes. Proportional abundances were calculated by dividing the number of species-specific sequence reads by the total number of unique sequence reads generated by each methodology. The relative abundance, as reported by BEI Resources, of each species found in HM-277D is also reported.
Reads were considered species-specific if they exhibited at least 97% identity over 97% of the sequence read length when compared to reference 16S rRNA genes, provided those reads only mapped to one species’ 16S rRNA gene. This approach excludes reads corresponding to conserved regions of the 16S rRNA gene, permitting unequivocal association of a given read with a species in HM-277D. Even though only 36.3% of the SEC-MC dataset mapped uniquely to individual 16S rRNA genes from HM-277D, reads identifying each species within the HM-277D community were found within the captured fragments. The SEC protocol successfully retrieved identifiable portions of the 16S rRNA genes from all 20 organisms within HM-277D regardless of their relative abundance in the mock community. Reads ranged from 239 *Bacteroides vulgatus*-specific counts to 27,799,733 *Escherichia coli*-specific counts (Table 3.3).

Since only 36% of SEC-MC mapped specifically to 16S rRNA genes, reads from this dataset were also mapped against the entire HM-277D mock community using CLC. This approach was taken to determine the fraction of reads clearly identifiable as part of this mock community and to characterize their distribution within the mock community. Sequence reads that shared 80% identity over 80% of the sequence read length were mapped without regard to whether they were found in more than one species within the mock community. The relaxed mapping parameters permitted identification of reads shared between organisms, including the conserved regions of the 16S rRNA gene. Using the relaxed parameters, 93% of the SEC-MC dataset could be unequivocally mapped to sequences within the genomes of the 20 organisms that are HM-277D (Figure 3.2).
The remaining 7% could not be assigned to a HM-277D genome sequence. These unassigned sequences were either too short, or the sequences did not match to a genome sequence with 80% identity. It is assumed that these sequences are artifacts of sequencing library preparation. Of the sequence reads that did map to the genome, the majority (56% of dataset) was associated with a variable or conserved region of the targeted loci – the 16S rRNA genes (Figure 3.2). In contrast to the low backgrounds observed during capture of the gltS sequence (Tables 3.1 and 3.2), 37% of the SEC-MC dataset mapped to other parts of the HM-277D mock community. Most of the apparent carryover (31% of the SEC-MC dataset) mapped to a single 8,007 bp plasmid (pSE-12228-03) found in *S. epidermidis* ATCC12228. The excessive amount of plasmid-derived sequences did not suggest a non-specific interaction; ATCC12228 carries six circular plasmids, four of which are under 8,100 bp, and only pSE-12228-03 appeared in this abundance in SEC-MC. Examination of pSE-12228-03 sequence revealed considerable complementarity between the biotinylated primer used to capture the 16S
rRNA gene and this plasmid that was not identified before attempting SEC. Fourteen of the 18 bases that make up the capture primer anneal perfectly to the plasmid, including the eight bases that comprise the 3´ end of the capture primer. It is assumed that the conditions for annealing the biotinylated primer during SEC allowed pSE-12228-03 to be captured and recovered during the protocol. To our knowledge, the copy number of pSE-12228-03 has not been characterized, and reasons for the number of pSE-12228-03-derived reads in SEC-MC was not investigated.

The remaining 6% of the SEC-MC dataset aligned to chromosomally encoded loci that were not found within a 16S rRNA gene or pSE-12228-03. For most of these sequences, there is no pattern to their distribution within the mock community, indicating they were non-specifically carried through to sequencing. However, almost 350,000 reads (approximately 0.34% of the recovered sequence reads) mapped to regions directly upstream of 16S rRNA genes. These are the only locations where the concentration and distribution of sequence reads argues against low levels of non-specific carryover. These sequences are presumably derived from long biotinylated DNA fragments created during linear amplification that were captured with the 16S rRNA genes, reflecting SEC’s ability to retrieve downstream sequence adjacent to the capture primer binding site.

Table 3.3 compares the reported abundance in HM-277D with a proportional abundance calculated from the number of species-specific read counts obtained from SEC-MC; the table is organized from the most to the least abundant member of the community. Figure 3.1 plots this information. The red symbols identify the values for reported abundance; there are four distinct levels of abundance present in HM-277D, each level differing by approximately an order of magnitude. Since the reported values for 18 of the 20 organisms in the mock community are given as either 0.02, 0.2, 2, or 20%, only six red symbols appear on the graph. The black
symbols correspond to the proportional abundance calculated from the species-specific sequence reads reported in Table 3.3 matched with their expected abundance as reported by BEI Resources. This plot generates four horizontal clusters of symbols, representing the four levels of abundance present in the mock community. Deviation from the reported abundances is apparent in the distribution of values for estimated abundance within each cluster. All estimated abundances were within an order of magnitude of their expected value with differences ranging from 1.5 to 20-fold. The mean difference between estimated and expected values was 4.9-fold over all twenty species in HM-277D.

**Defining the Composition of HM-277D through amplification of the V3-V5 region of the 16S rRNA gene**

Following the recommendation of the Jumpstart Consortium Human Microbiome Project Data Generation Working Group, HM-277D was also characterized using PCR amplification of the V3-V5 region of the 16S rRNA gene. Primers amplifying the V3-V5 region were combined with the DNA of HM-277D. Following amplification, the resulting PCR products were sequenced as described for the DNA captured by SEC. Sequencing the amplified V3-V5 region generated 171,983,220 raw reads, which after trimming yielded a dataset (designated PCR–MC) of 152,138,436 reads. Species-specific reads were mapped with CLC using the same mapping parameters applied to the SEC-MC dataset. Sequence reads were found that uniquely map to the V3-V5 region to the 16S rRNA genes of all 20 species in HM-277D. Read counts associated with individual species in PCR-MC are listed in Table 3.3.

Figure 3.3 plots estimated abundances (black) based on reads identified from sequencing the amplified V3-V5 regions while the red symbols represent the reported relative abundance. Read counts and the calculated relative abundances derived from the PCR-based characterization of HM-277D demonstrated greater variability than abundances calculated from SEC-MC. At
each level of abundance, at least one species was underestimated by 1000-fold or more. The problem of PCR biases in characterizing microbial environments has been well documented (Suzuki and Giovannoni 1996, Schloss, Gevers et al. 2011) and it appears to be reflected in the data obtained from V3-V5 amplification in this study. Some of the read counts from PCR-MC correlated with their reported abundances (6/20 organisms), but many (14/20) failed to adequately represent the reported levels of species’ DNA in HM-277D.

Figure 3.3 Reported and PCR-MC derived abundances of organisms within HM-277D. Reported abundances of organisms in HM-277D ranged from 0.02-26% and are shown in red. Abundance as determined by species-specific sequence reads from PCR-MC were plotted against the reported abundance (black).
Analyzing Sequence Data Obtained by SEC with EMIRGE

If SEC coupled to high-throughput sequencing is to be useful in defining the bacterial composition of an uncharacterized environmental sample, it will be necessary to identify alternative means of analyzing the resulting sequence data. As we have applied it, CLC Genomics Workbench would be of no use in mapping reads from novel experimental data; one would not have the luxury of knowing the composition of a community prior to analysis, as was the situation with the HM-277D mock community.

EMIRGE is a program that assembles 16S rRNA genes from short-read sequencing data using an expectation-maximization algorithm, and approximates relative abundances of assembled taxa (Miller, Baker et al. 2011). It has been applied to analyses of metagenomic sequence datasets, allowing investigators to catalogue species represented in the metagenome by assembling the 16S rRNA genes from sequence reads in the dataset (Miller, Baker et al. 2011). A SEC-derived dataset is similar to a conventional metagenomic dataset. It is composed of randomly generated short-read sequences, differing only in that 16 rRNA-specific reads are present in higher density compared to a metagenomic dataset.

To gain insight into how effective EMIRGE might be in describing SEC acquired sequencing data, the program was initially tested using a simulated dataset (designated simulated-staggered) containing five million error-free sequence reads; the relative abundance of species-specific 16S rRNA present in this simulation duplicates the distribution of read counts determined by CLC for SEC-MC and reported in Table 3.3. This simulated dataset was generated with wgsim to evaluate whether EMIRGE could accurately recreate full-length 16S rRNA gene sequences from an SEC-derived dataset. The analysis was independently replicated three times. EMIRGE assembled 77 full-length 16S rRNA gene sequences (average length of
1,455 ± 79 bases) that were then taxonomically classified using nucleotide BLAST (BLASTN) against the Silva (http://www.arb-silva.de/) database (SSUREF_111_NR). When classified, sequences were collapsed to genus-level taxa. Sixteen of the 17 genera in HM-277D were unequivocally identified in this analysis (Figure 3.4). EMIRGE was unable to recreate the *B. vulgatus* 16S rRNA gene, which was represented by only 29 unique sequence reads in the dataset. Estimates of relative abundance accurately reproduced the composition of the simulated-staggered data, (Figure 3.4), indicating the EMIRsGE output accurately defines the sequence read distribution within the sample dataset.

![Abundance percentages of genera from simulated-staggered dataset analyzed with CLC (orange) and EMIRGE (blue). EMIRGE Corrected (light blue) includes EMIRGE output that was corrected using BLAST against NCBI database and the RDP Classifier against the RDP database.](image)

Figure 3.4 Abundance percentages of genera from simulated-staggered dataset analyzed with CLC (orange) and EMIRGE (blue). EMIRGE Corrected (light blue) includes EMIRGE output that was corrected using BLAST against NCBI database and the RDP Classifier against the RDP database.
The EMIRGE output also assembled fourteen 16S rRNA gene sequences, representing seven identifiable genera not present in the dataset – an unacceptably high false positive rate. To verify this result, the false positive output sequences (the 16S rRNA sequence generated by EMIRGE) and the Silva database reference sequences that identified the false positives were re-classified again using a) BLASTN against the NCBI database and b) the naïve Bayesian classifier provided by the Ribosomal Database Project (RDP) (Wang, Garrity et al. 2007). Surprisingly, the results of these analyses did not match those found using the SILVA database, failing to associate any of the EMIRGE-assembled 16S rRNA genes with anything other than the species found in HM-277D. In each circumstance, it was determined that the SILVA reference sequence was misidentified in the database. For example, comparing the EMIRGE output to the Silva database identified a specific *Rhodococcus* sequence as a component of HM-277D, but a BLASTN search using this *Rhodococcus* sequence to query the NCBI database revealed a base for base match to an *Escherichia coli* 16S rRNA gene. Once the additional nucleotide BLASTN searches confirmed that a 16S rRNA sequence assembled by EMIRGE was incorrectly identified in our initial characterization, those sequences were excluded from the analysis. Following these corrections, EMIRGE only identified genera found within the HM-277D (Figure 3.4).

It was assumed that EMIRGE’s inability to recreate the full-length 16S rRNA sequence of *B. vulgatus* was related to the low number of *B. vulgatus*-specific sequence reads present in the simulated-staggered dataset, but it was unclear whether the failure was explained by the low numbers of sequence reads or the fact that the reads did not provide sufficient coverage to allow accurate assembly of the 16S rRNA gene of this genus. *D. radiodurans* presented the next lowest number of sequence reads (422 reads) in this dataset, and its genus was successfully assembled and identified. To ascertain the reasons for the failure to recreate the *B. vulgatus* 16S
rRNA gene, EMIRGE was tasked with assembling a series of 100-base, error-free, simulated datasets that varied in the number of \textit{B. vulgatus}-specific reads available. These analyses established that when using default parameters, EMIRGE needs at least 190 species-specific sequence reads that form a consensus length (the length of the 16S rRNA gene covered by at least one sequence read) greater than 500 bases to successfully assemble a 16S rRNA gene sequence. Reconstruction of the 16S rRNA gene was not possible when these minima were not met.

The SEC-MC dataset was too large to analyze in its entirety using EMIRGE. Instead, 5% of the dataset was randomly sampled using a custom script provided by Christopher Miller at the University of Colorado at Boulder (personal communication) prior to analysis. Three of these subsamples of SEC-MC were generated, and EMIRGE tasked to reassemble the 16S rRNA gene from each subsample independently. EMIRGE assembled 727 ± 4 16S rRNA gene sequences with an average length of 1329 ± 119 bases. BLASTN, performed against the Silva database (SSURef 111 NR), was used to determine taxonomic classification of the assembled sequences. Sequences were collapsed to genus-level classification (Figure 3.5). Using this data, EMIRGE identified 15 of the 17 genera present in two of the subsampled datasets and 14 of the 17 genera in the final subsample. A total of fifty genera not found in HM-277D were falsely identified from all three subsampled datasets. The existence of the incorrectly identified genera could not be confirmed when the assembled sequences were classified using BLASTN against the NCBI database. The EMIRGE program predicts relative abundance based on the number of reads used to reconstruct full-length 16S rRNA gene sequences. The relative abundances of the 15 genera recreated by EMIRGE reflected the relative abundances predicted by read count mappings according to CLC Genomics Workbench (Figure 3.5).
EMIRGE failed to assemble the 16S rRNA genes from *Bacteroides* and *Deinococcus* in each of the three subsamples derived from SEC-MC. The number of sequence reads connected to these genera was determined independently by mapping each subsample to their respective 16S rRNA genes with CLC Genomics Workbench (Table 3.4). On average, only 12 *Bacteroides*-specific and 141 *Deinococcus*-specific reads were found in each subsample. Based on our work with simulated datasets (see above), there were too few reads to support assembly by EMIRGE. In addition, EMIRGE failed to assemble the 16S rRNA gene from *Acinetobacter baumannii* in one of the three subsamples. Even though on average over 4,000 *Acinetobacter*-specific sequence reads were present in each subsample, the distribution of *Acinetobacter*-specific reads varied between subsamples, and coverage of the 16S rRNA gene differed. EMIRGE successfully assembled the *Acinetobacter* 16S rRNA gene from samples where
overlapping reads covered 500 bp of the gene (Table 3.4). In contrast, the reads in the third subsample formed a contiguous fragment that was only 150 bp long.

Table 3.4 Unique species-specific consensus lengths and total read counts of captured 16S rRNA genes from SEC-MC subsampled datasets.

<table>
<thead>
<tr>
<th>Species</th>
<th>Consensus length</th>
<th>Total read count</th>
</tr>
</thead>
<tbody>
<tr>
<td><em>Streptococcus mutans</em></td>
<td>1,503 ± 16</td>
<td>146,520 ± 635</td>
</tr>
<tr>
<td><em>Escherichia coli</em></td>
<td>1,497 ± 6</td>
<td>1,352,426 ± 1,050</td>
</tr>
<tr>
<td><em>Bacillus cereus</em></td>
<td>1,394 ± 0</td>
<td>31,200 ± 209</td>
</tr>
<tr>
<td><em>Propionibacterium acnes</em></td>
<td>1,385 ± 0</td>
<td>773 ± 48</td>
</tr>
<tr>
<td><em>Rhodobacter sphaeroides</em></td>
<td>1,334 ± 35</td>
<td>49,738 ± 40</td>
</tr>
<tr>
<td><em>Clostridium beijerinckii</em></td>
<td>1,342 ± 8</td>
<td>6,737 ± 160</td>
</tr>
<tr>
<td><em>Streptococcus agalactiae</em></td>
<td>1,354 ± 17</td>
<td>24,799 ± 28</td>
</tr>
<tr>
<td><em>Lactobacillus gasseri</em></td>
<td>1,220 ± 48</td>
<td>348 ± 35</td>
</tr>
<tr>
<td><em>Listeria monocytogenes</em></td>
<td>1,197 ± 3</td>
<td>2,329 ± 64</td>
</tr>
<tr>
<td><em>Pseudomonas aeruginosa</em></td>
<td>1,079 ± 189</td>
<td>7,072 ± 129</td>
</tr>
<tr>
<td><em>Enterococcus faecalis</em></td>
<td>1,029 ± 28</td>
<td>740 ± 7</td>
</tr>
<tr>
<td><em>Staphylococcus epidermidis</em></td>
<td>1,047 ± 1</td>
<td>104,936 ± 33</td>
</tr>
<tr>
<td><em>Staphylococcus aureus</em></td>
<td>1,027 ± 19</td>
<td>10,680 ± 127</td>
</tr>
<tr>
<td><em>Neisseria meningitides</em></td>
<td>978 ± 75</td>
<td>198 ± 8</td>
</tr>
<tr>
<td><em>Actinomyces odontolyticus</em></td>
<td>781 ± 0</td>
<td>6,380 ± 49</td>
</tr>
<tr>
<td><em>Helicobacter pylori</em></td>
<td>721 ± 51</td>
<td>1,726 ± 27</td>
</tr>
<tr>
<td><em>Streptococcus pneumoniae</em></td>
<td>769 ± 32</td>
<td>794 ± 24</td>
</tr>
<tr>
<td><em>Acinetobacter baumannii</em></td>
<td>674 ± 111</td>
<td>4,075 ± 16</td>
</tr>
<tr>
<td><em>Deinococcus radiodurans</em></td>
<td>425 ± 109</td>
<td>141 ± 9</td>
</tr>
<tr>
<td><em>Bacteroides vulgatus</em></td>
<td>417 ± 81</td>
<td>12 ± 3</td>
</tr>
</tbody>
</table>

Note: Read counts are the sum of the species-specific sequence reads found in the datasets. Consensus length is the length of the 16S rRNA gene reference covered by sequence reads. Values are the means ± standard deviations.

Since EMIRGE allows reconstruction of full-length 16S rRNA gene, this tool should provide better phylogenetic classification within the bacterial communities being characterized, as more of the 16S rRNA gene is available. As indicated in Figures 3.4 and 3.5, inferring genus-level distinctions within HM-277D and SEC-MC communities was easily accomplished with EMIRGE. Of the 17 genera represented in HM-277D, *Staphylococcus* and *Streptococcus* are represented by more than one species: there are three species of *Streptococcus* and two of
Staphylococcus. BLASTN against the NCBI database was used to classify the sequences generated by EMIRGE. The highest scores returned for these sequences identified 13 of the 17 species reconstructed by EMIRGE. Both Staphylococci were detected at the species level while two of the three Streptococci species were classified. The EMIRGE outputs for E. coli, B. cereus, E. faecalis, and S. pneumoniae failed to contain the highest score after BLASTN analysis against the NCBI database. While species level classification was determined for 13 of the 17 species identified in the EMIRGE output from SEC-MC, this method allows for distinct genera classification provided EMIRGE requirements are met.

**Discussion**

Two commonly used experimental approaches for defining bacterial diversity in a microbial community are to amplify and sequence a portion of the 16S rRNA gene (or some other conserved locus), or sequence the community’s metagenome and extract identifying gene sequences from that dataset. The approaches are not mutually exclusive, and both provide a means of identifying the species present. Assuming adequate sequencing depth, that is the average number of times a nucleotide appears in a sequence read (Sims, Sudbery et al. 2014), obtaining a metagenome provides a more comprehensive look at a community. However, there are many situations where simply defining a community through its 16S rRNA gene sequences is sufficient for an investigator’s purpose. SEC could provide a means of obtaining just the 16S rRNA genes from a bacterial community.

In this study, a method was developed to limit some PCR bias by using a single primer to target specific sequences from mixtures of DNA. Initially, we tested the protocol by targeting the gltS gene in purified E. coli genomic DNA. The results of this evaluation (Table 3.1) established that SEC recovered DNA fragments as large as 6,200 bases, did so with low
carryover of non-targeted genomic DNA, and could be applied to as little as 0.3 ng of purified DNA (Table 3.2). Variations between the amount of biotinylated product recovered from site 1 and 2 in Table 3.1 could be attributed to the shearing affects from the multiple washes required from the protocol. These results however, indicated that SEC could be used to recover any bacterial locus for which a capture primer could be designed. Extending these observations, we tested whether SEC could be used to facilitate characterization of a bacterial community by retrieving 16S rRNA genes from isolated DNA from a mock community prior to high-throughput sequencing. Capture with a biotinylated primer conveniently separated the 16S rRNA genes from non-targeted DNA, enriching the sample to be sequenced with these loci without a need for PCR amplification.

For these studies, we relied on the availability of DNA isolated from the bacterial mock community HM-277D. Since the composition of the HM-277D mock community is known, we were able to evaluate SEC for the ability to successfully identify the members of that community. Data analysis relied on mapping sequence reads to a representative 16S rRNA gene from each member of HM-277D, and recording the read counts. The CLC Genomics Workbench provided a convenient platform for performing these analyses; mapping parameters could be set that excluded sequence reads that mapped to more than one species and the software simplified determining the read counts.

Table 3.3 compares the effectiveness of SEC in enriching the sample to be sequenced with copies of the twenty 16S rRNA genes associated with HM-277D with a sample obtained following amplification of V3-V5 region of the 16S rRNA gene. While the SEC- and PCR-based methods generated sequence reads that identified the 20 species in HM-277D, the distribution of sequence reads among species differ dramatically when the two methods are
compared. SEC provided more species-specific sequence reads for the eleven species with proportional abundances of less than 0.2%. This effect is presumably a consequence of SEC retrieving more than a single region of the 16S rRNA gene, and as a result more species-specific reads are available to be sequenced.

When designing SEC, we assumed that the absence of PCR amplification prior to capture would result in a sequence read dataset where the distribution of species-specific reads would approximate the relative abundance of each species in the community being analyzed. As indicated in Figure 3.1 and Table 3.3, the read counts generated following SEC and high-throughput sequencing were within an order of magnitude to the reported abundances in HM-277D with differences ranging from 1.5-20 fold. PCR amplified V3-V5 regions of the 16S rRNA gene (PCR-MC) was less successful in predicting relative abundances. Four organisms within HM-277D had predicted abundances that differed by as much as 1,000-fold compared to their respective reported abundances (Table 3.3, Figure 3.3). (Jumpstart Consortium Human Microbiome Project Data Generation Working 2012) analyzed a similar mock community by PCR amplifying V3-V5 of the 16S rRNA gene and showed several members were either over- and underrepresented. These results mirror our results for the same amplified region.

Though SEC was more successful in predicting relative abundances of organisms in HM-277D than PCR amplification of V3-V5, possible biases could have been introduced during the library preparations for Illumina sequencing resulting in under and over amplification of members of HM-277D. At the time of sequencing, PCR amplification was a necessary step performed on the biotinylated product and could have introduced some biases. Aird et al. 2011 showed biases were introduced during the PCR step of Illumina library preparation steps. Currently, PCR-free library preparation kits are available and could eliminate any PCR biases
introduced during library preparation. Additionally, primers play an important role in amplification as primer selection could affect the amplification of templates (Klindworth, Pruesse et al. 2013). The biotinylated Uni1390R was complementary to 19 of the 20 organisms in HM-277D. A single mismatch was discovered in the 12th base from the 3’ end of biotinylated Uni1390R and the 16S rRNA genes of *Helicobacter pylori*. While internal mismatches found between primer and template can decrease amplification of the template (Sipos, Szekely et al. 2007), this was not observed for SEC-MC as the abundance percentage for mapped reads corresponding to *H. pylori* mirrored the reported abundance (Figure 3.1, Table 3.3). The ability to use a single primer provides an advantage to using SEC compared to PCR – as use of a single primer can decrease the potential primer-template mismatches that can introduce amplification biases. Additionally, PCR based methods for determination of bacterial diversity rely on successful amplification of bacterial 16S rRNA genes, and taxon-specific factors can negatively affect amplification. The “universal” primer binding sites are not well conserved in all taxa, resulting in poor amplification from some species (Jumpstart Consortium Human Microbiome Project Data Generation Working 2012). Furthermore, PCR requires the use of two primers and mismatches between primer pairs and templates can accumulate to result in unequal amplification (Klindworth, Pruesse et al. 2013). 16S rRNA gene sequences that amplify poorly are likely to be ignored and with a limited amount of product formed from lower abundant organisms, the result could be too unreliable to report.

Comparison of primers used to amplify the 16S rRNA gene were evaluated and compared with the biotinylated Uni1390R. Klindworth et al. 2013 provides a comprehensiva look at the bacterial coverage of common 16S rRNA primers used. Primers specific for the V6 region of the 16S rRNA gene used for pyrotags (Sogin, Morrison et al. 2006) poorly matched
bacteria with no mismatches. These primers, named S-D-Bact-0967-a-S-19 and S-D-Bact-1046-b-A-19, contained coverage of 60.4 and 53.5% of bacteria for the forward and reverse primers, respectively (Klindworth, Pruesse et al. 2013). Allowing for a single mismatch increases the coverage of bacteria to 80.9 and 96.6%, respectively (Klindworth, Pruesse et al. 2013). Another popular primer pair used to amplify the variable regions 1-8, S-D-Bact-0008-a-S-20 and S-*-Univ-1492-a-A-21, covered 67.4 and 28% of bacteria with no mismatches, respectively (Klindworth, Pruesse et al. 2013). The coverage of bacteria increases to 88.7 and 84.7% when allowing a single mismatch (Klindworth, Pruesse et al. 2013). Klindworth et al. 2013 also reported 90.7% coverage of bacteria for the forward primer used for amplification of V3-V5 (S-D-Bact-0337-a-S-20) with no mismatches; this number increases to 95.1 allowing for a single mismatch. The reverse primer (U926) could not be evaluated due to the presence of inosine bases in the primer. Coverage of the biotinylated Uni1390R were 69.2% for bacteria allowing for zero mismatches between template and primer and 79% allowing for a single mismatch (Klindworth, Pruesse et al. 2013). Due to the protocol of SEC, either a forward or reverse primer targeting bacteria may be used. Several primers evaluated by Klindworth et al. 2013 provided better bacterial coverage than the biotinylated capture primer used in this study. The primer designated S-D-Bact-0008-c-S-20 is specific for bacteria and covers 86.4% with no mismatches and 92.6% with a single mismatch. This primer may be better suited for capture of the 16S rRNA gene from bacteria and can be used as a biotinylated capture primer by modifying the 5´ end to contain a biotin molecule.

Since the use of the CLC Genomics Workbench in classifying sequences requires knowledge of the sample composition prior to analysis, we explored the use of EMIRGE. EMIRGE, a program that takes advantage of SEC’s ability to capture eight of the nine variable
regions of the 16S rRNA gene, recreated nearly full-length 16S rRNA gene sequences from SEC-generated data. Working with a simulated dataset, EMIRGE reproduced the composition of a staggered bacterial community containing 20 species, successfully reconstructing 16 of the 17 genera, failing to reconstruct the 16S rRNA gene from Bacteroides, the least abundant member present. Unfortunately, analysis with EMIRGE is computationally intensive, requiring extended runtime and memory. As a result, SEC-MC was subsampled in triplicate at 5% and analyzed with EMIRGE. Analysis of EMIRGE assembled sequences using BLASTN against the NCBI database resulted in genus-level classification. Assembled sequences for 15 of the 17 genera were present in two of the subsamples while 14 of the 17 genera were present in the final subsample. Species-level classification was not accurately obtained as multiple assembled sequences from BLASTN analysis matched to species not found in HM-277D. A possible explanation for this could be the way in which EMIRGE reconstructs full-length sequences. EMIRGE maps sequence reads against Silva reference sequences to calculate the probability of each read belonging to the reference sequence. When a certain probability is calculated, the sequence read from the dataset replaces the reference sequence. Areas along the reference sequence that do not contain reads or enough coverage will result in the reference sequence being used instead of sequence reads for reconstruction. As a result, assembled full-length 16S rRNA genes include both sequence reads from the dataset and reference sequence that affect the ability to classify to species-level.

While classification of species was not possible, the predicted abundances of the assembled EMIRGE sequences were similar to the predicted amounts using CLC Genomics Workbench (Figure 3.5). This provides evidence that EMIRGE determines relative abundance based on species-specific sequence reads for each organism within a given dataset; any potential
biases within a dataset will be reflected in the output from EMIRGE. Genera of 16S rRNA genes not reconstructed either contained too few species-specific reads, or those reads did not cover enough of the locus to permit reconstruction of their 16S rRNA genes. Making more of the dataset available for analysis should increase the probability of detecting these genera and gives insight into the limitations of the program.

The effectiveness of SEC in capturing 16S rRNA genes and analysis with EMIRGE resulted in the assembly of full-length 16S rRNA genes. Yarza, Yilmaz et al. 2014 report that full-length 16S rRNA genes result in more accurate taxonomic classifications for bacteria when compared to fragmented variable regions of the 16S rRNA gene showcasing the usefulness of SEC. Full-length 16S rRNA genes also define species richness and evenness more accurately than PCR amplified variable regions (Youssef, Sheik et al. 2009, Jumpstart Consortium Human Microbiome Project Data Generation Working 2012, Yarza, Yilmaz et al. 2014), which agree with the results obtained in this work. While combinations of multiple variable regions improve taxonomic classification compared to single variable regions, it does not surpass results achieved with full-length 16S rRNA genes (Jumpstart Consortium Human Microbiome Project Data Generation Working 2012, Yarza, Yilmaz et al. 2014) validating an advantage to using SEC over PCR based methods.

When coupled with high-throughput sequencing, the SEC protocol facilitates the isolation, identification, and quantitation of specific DNA sequences within a complex mixture. It is a protocol that will allow the investigator to retrieve the proverbial “needle in a haystack.” The capacity to selectively target any DNA sequence with a single capture primer is perhaps the most useful feature of the methodology. One only needs to know the sequence of a short fragment of DNA adjacent to what is to be retrieved, and the protocol can be implemented.
Knowledge of the sequence 3’ to this known short sequence is unnecessary as downstream sequence is copied. In addition, SEC could eliminate some PCR biases introduced as a result of primer-template mismatches by using a single primer. SEC usage on the capture of 16S rRNA genes can provide an alternative means of obtaining the microbial community structure without the need for PCR amplification.
CHAPTER 4.
DETERMINATION OF THE MICROBIAL DIVERSITY WITHIN THE EASTERN OYSTER, *CRASSOSTREA VIRGINIA* BEFORE AND AFTER DEPURATION TREATMENT, AND COMPARISON WITH SURROUNDING SEDIMENT AND SEAWATER

Introduction

Chapter 3 was an investigation into how Selected Enrichment through Capture (SEC) may be used to target and retrieve specific loci from mixtures of genomic DNA. The method was tested on a single locus from *E. coli*, the *gltS* gene. Recovery of the entire *gltS*, along with 5,000 bases of adjacent DNA was performed efficiently. Then, SEC was tested on a more complex sample, targeting a conserved locus found in an artificial bacterial mock community composed of 20 species with varied abundances. The captured material was sequenced using an Illumina HiSeq 2000 platform to determine sequence identity. SEC could effectively capture 16S rRNA genes from each of the twenty organisms regardless of a species’ relative abundance in the sample. The results demonstrated that SEC was effective in qualitatively defining bacterial diversity in this relatively simple bacterial community.

Prior knowledge of the bacterial diversity in an environmental community is not a realistic expectation, and methods like PCR amplicon and metagenomic sequencing are routinely used to analyze these communities of unknown composition. Previous discussion defined problems associated with PCR amplicon sequencing when attempting to determine bacterial diversity and will not be focused on here. Metagenomic sequencing, defined as sequencing all DNA extracted from an environmental sample, provides a means of analyzing the microbial diversity without PCR amplification, offering a potentially bias-free look at the microbial diversity in environmental samples, but this type of analysis also has limitations that must be considered before its undertaken. Unlike the artificial bacterial community analyzed in Chapter
3, environmental samples are not composed solely of bacteria; they include eukaryotes as well. The presence of eukaryotes presents a significant problem for those interested in examining the bacterial components of the environments they are located in. In general, microbial eukaryotes have a larger genome and disproportionally contribute to the pool of DNA sequenced during metagenomic analysis (Hou and Lin 2009). From a practical point of view, this means that greater sequencing depth is necessary to define prokaryotic community structure. The probability that species-specific sequence reads appear in the dataset is likely proportional to the relative abundance of that species’ genomic DNA within the mixture being sequenced, and the depth of sequencing during high-throughput metagenomic analysis. Since most bacteria have two to three orders of magnitude less DNA per cell when compared to microbial eukaryotes (Hou and Lin 2009), more eukaryotic DNA will be sequenced at all sequencing depths. Depending on sequencing depth, low abundance bacteria may be underrepresented or missed entirely (Sogin, Morrison et al. 2006). For example, it has been proven difficult to sequence the metagenome from the hindgut of lower termites due to the presence of protozoa. As a consequence, it was necessary to use higher termites when characterizing this environment with metagenomic sequencing because their hindguts do not contain protozoa and produce their own cellulases and hemicellulases (Warnecke, Luginbuhl et al. 2007). The absence of protozoa allowed for more of the prokaryotic genomes to be sequenced at greater depth. Physically removing eukaryotes by flow cytometry prior to DNA isolation (Cuvelier, Allen et al. 2010, Vaulot, Lepère et al. 2012) is a way to avoid this problem, but this approach can disrupt the microbial community structure, and may bias predictions concerning bacterial diversity.

In this study, SEC was evaluated as a potential solution to determining the bacterial diversity in an environment present with eukaryotes. The environments chosen for this study
contained increased complexity by having more species relative to the artificial bacterial mock community used in Chapter 3. Three distinct, but related, marine environments were examined: cultivated Eastern oysters (*Crassostrea virginica*) from an oyster cage in Grand Isle, Louisiana, seawater surrounding the oyster cages, and sediment below the oyster cage.

The Eastern oyster, *C. virginica* is widely distributed along the Atlantic coast and Gulf of Mexico (Banks 2007). It serves as a popular food source in Louisiana who consume more than 750,000 bushels of oysters per year (Banks 2007). In addition, Louisiana alone is responsible for 42% of all the oysters harvested in the United States of America and is one of the few states that market oysters year round (Banks 2007). *C. virginica* is also important ecologically for its ability to create reefs that can aid in the species diversity of the surrounding ecosystem (Banks 2007) and help reduce shoreline erosion (Piazza, Banks et al. 2005). Furthermore, *C. virginica* has been implicated as a possible source for bioremediation of estuaries due to their ability to filter seawater (Banks 2007). *C. virginica* were found to reduce total suspended solids in transplanted oyster reefs (Nelson, Leonard et al. 2004). A related species, *C. gigas*, was proposed for bioremediation purposes in fish farms for its ability to remove leftover food and waste improving water quality (Silva, Yanez et al. 2012).

Thus far, no study has attempted to determine the microbiome of whole oysters using high-throughput sequencing.


The principle objective of this work was to determine whether SEC could be used to retrieve bacterial 16S rRNA genes from these environmental samples in the presence of simultaneously isolated eukaryotic DNA, comparing those results with previous studies of microbial diversity in similar environments. SEC was used to: 1) determine the bacterial diversity of the seawater and sediment surrounding oyster cages, 2) define the bacterial microbiome of C. virginica, 3) investigate whether the bacteria identified in the seawater and
sediment reflect the microbiome of *C. virginica*, and 4) monitor changes in the microbiome of *C. virginica* after depuration treatment.

**Results**

**SEC on Seawater Surrounding Oyster Cages**

Seawater collected from Grand Isle, LA was filtered and SEC was performed on extracted DNA, targeting the 16S rRNA gene. Recovered material was sequenced using the Illumina HiSeq 2000 platform. The dataset, designated SEC-Seawater, was trimmed with CLC and resulted in 63,670,589 reads. This was subsampled five times with each subsample representing 10% of the dataset using a custom script (Miller 2013), and analyzed with EMIRGE. Subsampling the dataset was necessary due to computational constraints using the program EMIRGE. The EMIRGE output recreated 141 ± 7 16S rRNA gene sequences with an average length of 1,050 ± 50 bases.

The bacterial diversity was dispersed over the following phyla: Bacteroidetes (43.2 ± 4.5%), Proteobacteria (35.5 ± 5.8%), Cyanobacteria (18.7 ± 2.3%), and Actinobacteria (2.6 ± 1.7%) (Figure 4.1). The distribution of bacterial phyla observed was similar to that found in a recent study of Northern Gulf of Mexico seawater unaffected by the DHOS obtained near St. George Island, Florida (Figure 4.1) (Newton, Huse et al. 2013). St. George Island is located at similar latitude as Grand Isle. Though Newton et al. (2013) sampled seawater at different months at St. George Island, the data used for comparison in Figure 4.1 and Table 4.1 was obtained during June at St. George Island, the closest timing with the date of sampling for this study. The St. George Island data is derived from pyrosequencing. The pyrotag dataset only classified 70% of the sequence reads obtained in that study, and percentages reported in Figure 4.1 do not include the unclassified reads reported. The majority of the pyrotags identified
bacteria associated with Proteobacteria (48%), Cyanobacteria (27%), and Bacteroidetes (25%) (Newton, Huse et al. 2013); similar to results obtained for at the Grand Isle site.

![Figure 4.1 Comparison of bacterial phyla identified from Grand Isle, LA seawater through SEC (blue) and St. George Island, FL seawater through pyrotags (red) from Newton et al. (2013). Values for SEC (blue) are averages ± standard deviation of bacterial phyla identified through EMIRGE.](image)

Table 4.1 Phylogenetic composition of bacteria found Grand Isle, Louisiana and St. George Island, Florida seawater.

<table>
<thead>
<tr>
<th>Phylum</th>
<th>Class</th>
<th>Family</th>
<th>Grand Isle</th>
<th>St. George Island*</th>
</tr>
</thead>
<tbody>
<tr>
<td>Actinobacteria</td>
<td>Actinobacteria</td>
<td>Acidimicrobiaceae</td>
<td>0.31 ± 0.33</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Microbacteriaceae</td>
<td>0.60 ± 0.66</td>
<td></td>
</tr>
<tr>
<td>Bacteroidetes</td>
<td>Flavobacteria</td>
<td>Cryomorphaceae</td>
<td>0.19 ± 0.18</td>
<td>2</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Flavobacteriaceae</td>
<td>11.67 ± 3.53</td>
<td>15</td>
</tr>
<tr>
<td></td>
<td>Sphingobacteria</td>
<td>Chitinophagaceae</td>
<td></td>
<td>1</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Saprospiraceae</td>
<td>0.56 ± 0.21</td>
<td>1</td>
</tr>
<tr>
<td>Cyanobacteria</td>
<td>Chroococcales</td>
<td>Chamaesiphonaceae</td>
<td>4.51 ± 0.78</td>
<td>20</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Rivulariaceae</td>
<td>0.68 ± 0.33</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Prochlorales</td>
<td>Prochlorococcaceae</td>
<td>0.36 ± 0.33</td>
<td></td>
</tr>
<tr>
<td>Proteobacteria</td>
<td>α-Proteobacteria</td>
<td>Rhodobacteraceae</td>
<td></td>
<td>8</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Rhodospirillaceae</td>
<td></td>
<td>2</td>
</tr>
<tr>
<td></td>
<td>β-Proteobacteria</td>
<td>SAR 11 cluster</td>
<td></td>
<td>6</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Comamonadaceae</td>
<td></td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>δ-Proteobacteria</td>
<td>Bacteriovoracaceae</td>
<td>0.58 ± 0.10</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Desulfovibrionaceae</td>
<td>0.18 ± 0.10</td>
<td></td>
</tr>
</tbody>
</table>
Bacterial diversity from EMIRGE analyzed SEC-Seawater revealed seventeen bacterial families in all five subsamples with a majority of the sequence reads belonging to

*Flavobacteriaceae* (11.7 ± 3.5%) followed by *Pseudoalteromonadaceae* (5.2 ± 2.7%), *Chamaesiphonaceae* (4.5 ± 0.8%), and *Alteromonadaceae* (2.1 ± 1.2). The remaining 13 families were identified through the assembled sequences and combined; these families each comprised less than 1% of the sequence reads in SEC-Seawater (Table 4.1). Of the seventeen bacterial families found at the St. George Island site, six were also identified in the SEC-Seawater dataset (Table 4.1). Families from Bacteroidetes were identified in both Grand Isle and St. George Island seawater suggests these families may be widespread in the Gulf of Mexico. Particularly members of *Flavobacteriaceae*, which comprised greater than 11% of the classified sequence reads from both studies. These results provide evidence that SEC can be used to capture 16S rRNA genes from an environmental sample and provide family level classifications of the bacterial diversity found in that environment.
SEC Retrieval of 16S rRNA Genes from Sediment

Five grams of sediment were combined with equal volumes of PBS and agitated with a vortex mixer to release cells from particles. Aliquots were used for DNA extractions. SEC was performed on extracted DNA and sequenced using the Illumina HiSeq 2000 platform.

As described previously, the dataset, designated SEC-Sediment, was trimmed with CLC and resulted in 49,865,188 reads. This was subsampled five times with a custom script (Miller 2013) with each subsample comprising of 10% of the dataset. EMIRGE was used to analyze the subsampled SEC-Sediment datasets resulting in 337 ± 31 16S rRNA gene sequences with an average length of 1,200 ± 50 bases assembled. The bacterial diversity included six phyla (Figure 4.2, blue bars) with a majority of sequence reads belonging to Proteobacteria (69 ± 26%), followed by Firmicutes (23.5 ± 21%), Fusobacteria (3.2 ± 7%), Bacteroidetes (3.3 ± 1.9%), Chlorobi (0.6 ± 0.3%), and Verrucomicrobia (0.5 ± 0.3%). Clone libraries (Mills, Hunter et al. 2008) derived from DNA isolated from marine sediment near St. George Island, Florida during the months of March and May, also detected Proteobacteria, Actinobacteria, Chloroflexi, and Firmicutes. More recently a study of the microbial diversity from St. George Island, Florida sediment using pyrotags, found Proteobacteria, Bacteroidetes, Planctomycetes, and Actinobacteria (Figure 4.2, red bars) (Newton, Huse et al. 2013), but did not report Firmicutes, Fusobacteria, Chlorobi, or Verrucomicrobia. These more recent results are presented in Figure 4.2 and represent only 55% of the sequence reads; the remaining 45% were unclassified and are not reported.
Thirty-one bacterial families were identified from SEC-Sediment. The vast majority of sequence reads from assembled sequences belonged to *Desulfovibrioaceae* (23 ± 18%) followed by *Desulfobulbaceae* (19 ± 9.5%), *Streptococcaceae* (17 ± 20), *Geobacteraceae* (6.0 ± 3.0%), and *Alteromonadaceae* (4.1 ± 2.7%). All remaining families contained sequence reads that were present less than 4% each of the relative abundance in the sequencing data (Table 4.2).

Comparing the bacterial families identified from this study (Grand Isle) to sediment from St. George Island, FL (Newton, Huse et al. 2013) identifies five bacterial families shared between the two sites, *Cytophagaceae*, *Flammeovirgaceae*, *Saprospiraceae*, *Desulfovibrioaceae*, and *Thioalkalivirgaceae* (Table 4.2). The bacterial family *Saprospiraceae* was also found in seawater from Grand Isle, LA and St. George Island (Table 4.1). Twenty-six bacterial families identified in Grand Isle sediment were not reported in St. George Island sediment. Several sulfur and sulfate-reducing bacterial families, *Desulfovibrioaceae*, *Desulfobulbaceae*, and
Desulfuromonadaceae were identified in Grand Isle sediment and has not been previously reported in Louisiana estuary sediments. Members from Geobacteraceae, identified as the genera Geoalkalibacter and Geobacter, are known iron-reducers (Castro, Williams et al. 2000, Zavarzina, Kolganova et al. 2006). EMIRGE also assembled a single sequence belonging to the bacterial family Nitrospinaceae, genus Nitrospina, known to oxidize nitrite (Watson and Waterbury 1971). The phototrophic Marichromatium (Chromatiaceae) are purple sulfur γ-Proteobacteria (Shivali, Ramana et al. 2011), the methan-oxidizer, Methylomonas (Methylococcaceae) (Sorokin, Jones et al. 2000), and a methylo-troph, Methylomonas (Ectothiorhodospiraceae) (Sorokin, Trotsenko et al. 2007) were also identified.

Table 4.2 Phylogenetic composition of bacteria found Grand Isle, Louisiana and St. George Island, Florida sediment.

<table>
<thead>
<tr>
<th>Phylum</th>
<th>Class</th>
<th>Family</th>
<th>Grand Isle</th>
<th>St. George Island*</th>
</tr>
</thead>
<tbody>
<tr>
<td>Acidobacteria</td>
<td>Acidobacteria</td>
<td>Acidobacteriaceae</td>
<td>2</td>
<td></td>
</tr>
<tr>
<td>Actinobacteria</td>
<td>Actinobacteria</td>
<td>Iamiaceae</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>Bacteroidetes</td>
<td>Bacteroidia</td>
<td>Bacteroidaceae</td>
<td>0.16 ± 0.13</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Prolibacteriaecae</td>
<td>0.23 ± 0.07</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Bacteroidetes Order II</td>
<td>Rhodothermaceae</td>
<td>3</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Cytophagia</td>
<td>Cytophagaceae</td>
<td>0.35 ± 0.19</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Flavmeovirgaceae</td>
<td>0.57 ± 0.65</td>
<td>3</td>
</tr>
<tr>
<td>Flavobacteria</td>
<td></td>
<td>Cryomorphaceae</td>
<td>1.19 ± 0.88</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Flavobacteriaceae</td>
<td>10</td>
<td></td>
</tr>
<tr>
<td>Sphingobacteria</td>
<td></td>
<td>Saprospriraceae</td>
<td>1.07 ± 1.08</td>
<td>7.5</td>
</tr>
<tr>
<td>Chlorobi</td>
<td>Ignivibacteria</td>
<td>Ignivibacteriaceae</td>
<td>0.66 ± 0.16</td>
<td></td>
</tr>
<tr>
<td>Chloroflexi</td>
<td>Anaerolineae</td>
<td>Anaerolinaceae</td>
<td>0.5</td>
<td></td>
</tr>
<tr>
<td>Firmicutes</td>
<td>Bacilli</td>
<td>Bacillaceae</td>
<td>0.14 ± 0.08</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Streptococcaceae</td>
<td>17.14 ± 20.80</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Clostridia</td>
<td>Clostridaceae</td>
<td>3.79 ± 4.86</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Eubacteriaceae</td>
<td>0.36 ± 0.36</td>
<td></td>
</tr>
</tbody>
</table>
(Table 4.2 continued)

<table>
<thead>
<tr>
<th>Phylum</th>
<th>Class</th>
<th>Family</th>
<th>Grand Isle</th>
<th>St. George Island*</th>
</tr>
</thead>
<tbody>
<tr>
<td>Firmicutes</td>
<td>Clostridia</td>
<td>Gracilibacteraceae</td>
<td>0.47 ± 0.38</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Thermoanaerobacteria</td>
<td>Thermoanaerobacteraceae</td>
<td>1.61 ± 0.53</td>
<td></td>
</tr>
<tr>
<td>Fusobacteria</td>
<td>Fusobacteria</td>
<td>Fusobacteriaceae</td>
<td>5.11 ± 8.55</td>
<td></td>
</tr>
<tr>
<td>Planctomycetes</td>
<td>Planctomycetia</td>
<td>Planctomycetaceae</td>
<td>7</td>
<td></td>
</tr>
<tr>
<td>α-Proteobacteria</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Desulfo bacteriae</td>
<td>Desulfo bacteriae</td>
<td>23.18 ± 17.94</td>
<td>2</td>
</tr>
<tr>
<td></td>
<td>Desulfobulbaceae</td>
<td>Desulfobulbaceae</td>
<td>19.35 ± 9.51</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Desulfuromonadaceae</td>
<td>Desulfuromonadaceae</td>
<td>2.80 ± 2.79</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Geobacteraceae</td>
<td>Geobacteraceae</td>
<td>6.06 ± 2.99</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Nitrospinaceae</td>
<td>Nitrospinaceae</td>
<td>1.65 ± 0.69</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Pelobacteraceae</td>
<td>Pelobacteraceae</td>
<td>0.61 ± 0.14</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Syntrophaceae</td>
<td>Syntrophaceae</td>
<td>2.12 ± 1.86</td>
<td></td>
</tr>
<tr>
<td>Proteobacteria</td>
<td>δ-Proteobacteria</td>
<td>Alcanivoraceae</td>
<td>3.65 ± 3.90</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Alteromonadaceae</td>
<td>Alteromonadaceae</td>
<td>4.14 ± 2.69</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Chromatiaceae</td>
<td>Chromatiaceae</td>
<td>0.38 ± 0.35</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Ectothiorhodospiraceae</td>
<td>Ectothiorhodospiraceae</td>
<td>1.87 ± 2.66</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Enterobacteriaceae</td>
<td>Enterobacteriaceae</td>
<td>2.10 ± 2.01</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Methyllococcaceae</td>
<td>Methyllococcaceae</td>
<td>0.35 ± 0.15</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Oceanospirillaceae</td>
<td>Oceanospirillaceae</td>
<td>0.23 ± 0.22</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Pseudomonadaceae</td>
<td>Pseudomonadaceae</td>
<td>0.24 ± 0.26</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Thioalkalispiraceae</td>
<td>Thioalkalispiraceae</td>
<td>0.33 ± 0.12</td>
<td>15</td>
</tr>
<tr>
<td>Verrucomicrobia</td>
<td>Opitutae</td>
<td>Opitutae</td>
<td>0.27 ± 0.21</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Verrucomicrobiae</td>
<td>Verrucomicrobiae</td>
<td>0.33 ± 0.23</td>
<td></td>
</tr>
</tbody>
</table>

Note: Values for Grand Isle sediment are shown as percentages and are averages ± standard deviation. Percentages were calculated by EMIRGE based on the number of sequence reads attributing to each assembled sequence found to be bacterial and were present at least three subsamples. *St. George Island sediment is data from the June sampling from Newton, Huse et al. (2013).

SEC Capture of the Microbiome of C. virginica

Retrieval of 16S rRNA genes from seawater and sediment demonstrated that SEC could be used with environmental samples. To further test the SEC protocol, an attempt was made to
define the microbiome of the Eastern oyster, *C. virginica* by capturing 16S rRNA genes from DNA that included the oyster’s genome; a system that would contain an excess of eukaryotic DNA.

Whole oysters were harvested from the Louisiana Sea Grant oyster hatchery near Grand Isle, LA. Each oyster was washed to remove sediment on the outside of the shells. Whole oysters were then shucked and homogenized. DNA extractions were also performed on aliquots of the whole oyster homogenates. Following DNA extraction, SEC was performed and captured material was sequenced using an Illumina HiSeq 2000. The dataset, designated SEC-Oyster, was trimmed with CLC and resulted in 43,028,970 reads.

SEC-Oyster was subsampled five times each at 10% of the dataset and analyzed with EMIRGE. The EMIRGE output revealed 202 ± 34 16S rRNA gene sequences with an average length of 590 ± 42 bases. Only two bacterial phyla were identified, Proteobacteria and Firmicutes (Figure 4.3). 89.9 ± 2.2% of the sequence reads were attributed to Proteobacteria, of which the γ-Proteobacteria and α-Proteobacteria accounted for 87.7 and 2.2% of the sequence reads, respectively. All Firmicutes were classified as Bacilli. Family level classification is presented in Table 4.3. Seven bacterial families were identified in the whole oyster homogenate; the most abundant at 85.8 ± 1.9% of the sequence reads was *Enterobacteriaceae*, followed by *Streptococcaceae* (8.9 ± 2.4%), and *Rhodobacteraceae* (2.1 ± 1.7%). Families *Bacillaceae*, *Staphylococcaceae*, *Enterococcaceae*, and *Vibrionaceae* all contained sequence reads with abundances of less than 0.6 percent each (Table 4.3).
Figure 4.3 Comparison of bacterial phyla identified from whole oyster homogenate from Grand Isle, LA through SEC (blue) and stomach and gut phyla from Hackberry Bay, LA through pyrotags (red) with data from King et al. (2012). Values are the averages ± standard deviation.

Table 4.3 Phylogenetic composition of bacteria identified through SEC capture from extracted DNA from whole oyster homogenate from Grand Isle, Louisiana.

<table>
<thead>
<tr>
<th>Phylum</th>
<th>Class</th>
<th>Family</th>
<th>Oyster</th>
</tr>
</thead>
<tbody>
<tr>
<td>Firmicutes</td>
<td>Bacilli</td>
<td>Bacillaceae</td>
<td>0.55 ± 0.34</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Staphylococcaceae</td>
<td>0.52 ± 0.15</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Enterococcaceae</td>
<td>0.22 ± 0.15</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Streptococcaceae</td>
<td>8.93 ± 2.36</td>
</tr>
<tr>
<td>Proteobacteria</td>
<td>α-Proteobacteria</td>
<td>Rhodobacteraceae</td>
<td>2.13 ± 1.68</td>
</tr>
<tr>
<td></td>
<td>γ-Proteobacteria</td>
<td>Enterobacteriaceae</td>
<td>85.81 ± 1.87</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Vibrionaceae</td>
<td>0.53 ± 0.60</td>
</tr>
</tbody>
</table>

Note: Values for Oyster are shown as percentages and are averages ± standard deviation. Percentages were calculated by EMIRGE based on the number of sequence reads attributing to each assembled sequence found to be bacterial and were present at least three subsamples.

A study in 2012 was performed that investigated the prokaryotic diversity within the stomach and gut of C. virginica through V4 iTags from two sites in Louisiana, Cailou Lake and Hackberry Bay during the month of August (King, Judd et al. 2012). In this study, the oyster’s stomach and gut were removed and the microbiome evaluated. The V4 iTags generated a
different community than described using whole oyster homogenates. King et al. (2012) identified eight bacterial and two archaeal phyla in the stomach and gut of oysters harvested from Hackberry Bay, located in the same estuary as our Grand Isle oysters (Figure 4.3, red bars). Only two phylum level classifications were made with the whole oysters in this SEC facilitated study (Figure 4.3, blue bars). SEC-Oyster did not contain assembled sequences from Actinobacteria, Mollicutes, Chloroflexi, Planctomycetes, or Verrucomicrobia, which accounted for 60.6% of OTUs from King et al. (2012). Additionally, 15% of the pyrotag dataset could not be classified into phyla and are not reported in Figure 4.3.

**Comparison of Microbial Diversity in Oyster and Surrounding Sediment and Seawater**

The question of whether the microbiome of *C. virginica* would most reflect the surrounding seawater or sediment was investigated by comparing EMIRGE-derived families. Oysters are filter feeders and gain nutrients through ingestion of surrounding seawater and sediment (Berg 1986) and oysters from the Grand Isle site were grown in cages that rested on the sediment. Figure 4.4 is a Venn diagram of bacterial families shared between the seawater, sediment, and oyster. Only the family *Enterobacteriaceae* was found in all three environments.

The microbiome of oysters harvested at Grand Isle shared more similarity with the sediment than seawater with three bacterial families – *Bacillaceae*, *Streptococcaceae* and *Enterobacteriaceae* – identified in both environments compared to just one with seawater (Figure 4.4).
SEC Retrieval of 16S rRNA Genes from the Microbiome of *C. virginica* After Depuration Treatment

Harmful bacteria, including *V. vulnificus* and *V. chloreae*, can inhabit *C. virginica* (Tamplin and Capers 1992, Murphree and Tamplin 1995, Motes, DePaola et al. 1998, Campbell and Wright 2003, Zimmerman, DePaola et al. 2007, Chae, Cheney et al. 2009). Depuration is a Food and Drug Administration (FDA) approved method to remove these pathogenic bacteria from *C. virginica* (Program 2009). Depuration involves placing oysters in holding tanks with artificial seawater that is continuously filtered and UV sterilized. Oysters will ingest the treated seawater purging transient and potentially harmful bacteria. The effect of depuration treatment on the microbiome of *C. virginica* has never been thoroughly investigated. In this study, we provide a first look at how depuration treatment affects the bacterial flora of the oyster.

Eight oysters to be depurated were harvested at the same time and location as control oysters described in Figure 4.3 and Table 4.3. Oysters were placed in a depuration tank at a salinity matching the harvest site. After one week, three treated oysters were removed, shucked,
and homogenized. SEC was performed and captured material was sequenced using an Illumina HiSeq 2000. The dataset, designated SEC-DepuratedOyster, was trimmed with CLC and resulted in 59,078,237 reads. This was subsampled five times with a custom script (Miller 2013) with each subsample containing 10% of the dataset. The subsampled datasets were analyzed with EMIRGE. EMIRGE analysis showed a decrease in total number of assembled sequences from 202 ± 34 for control oysters to 45 ± 19 after depuration despite SEC-DepuratedOyster containing 1.4 times more sequence reads than SEC-Oyster. Assembled sequences contained an average length of 755 ± 227 bases.

As observed with the untreated oysters, a majority of the sequence reads belonged to Proteobacteria which accounted for 100% of the sequence reads from assembled sequences (Figure 4.5). Two bacterial families were identified in the depurated oyster, *Vibrionaceae* and *Enterobacteriaceae* (Table 4.4). Assembled sequences fell into two genera, *Vibrio* and *Escherichia*. *Vibrio* dominated the percentage of sequence reads (81 ± 5%) followed by *Escherichia* (8.2 ± 5.9%).

![Figure 4.5 Bacterial phyla composition of the microbiome of oyster (purple) and depurated oyster (blue). Values are the averages ± standard deviation.](image-url)
Table 4.4 Phylogenetic composition of bacteria identified through SEC from oysters before and after depuration treatment.

<table>
<thead>
<tr>
<th>Phylum</th>
<th>Class</th>
<th>Family</th>
<th>Oyster</th>
<th>Depurated Oyster</th>
</tr>
</thead>
<tbody>
<tr>
<td>Firmicutes</td>
<td>Bacilli</td>
<td>Bacillaceae</td>
<td>0.55 ± 0.34</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Staphylococcaceae</td>
<td>0.52 ± 0.15</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Enterococcaceae</td>
<td>0.22 ± 0.15</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Streptococcaceae</td>
<td>8.93 ± 2.36</td>
<td></td>
</tr>
<tr>
<td>Proteobacteria</td>
<td>α-Proteobacteria</td>
<td>Rhodobacteraceae</td>
<td>2.13 ± 1.68</td>
<td></td>
</tr>
<tr>
<td></td>
<td>γ-Proteobacteria</td>
<td>Enterobacteriaceae</td>
<td>85.81 ± 1.87</td>
<td>8.20 ± 5.87</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Vibrionaceae</td>
<td>0.53 ± 0.60</td>
<td>80.69 ± 5.06</td>
</tr>
</tbody>
</table>

Note: Values for Oyster and Depurated Oyster are shown as percentages calculated by EMIRGE and are averages ± standard deviation. Percentages were calculated by EMIRGE based on the number of sequence reads attributing to each assembled sequence found to be bacterial and were present at least three subsamples.

Two bacterial families identified in *C. virginica* following depuration treatment were also found in the oyster prior to depuration, indicating that these microorganisms were not completely removed the treatment. *Escherichia* exhibited a large decrease in percentage of sequence reads from 86% before depuration and decreasing to 8% following depuration. These results suggest that depuration allows *C. virginica* to purge a large portion of *Escherichia* organisms, presumable those found in the digestive tract. The largest difference was observed in percentage of sequence reads for the *Vibrio*, which was 0.53% before depuration and increased to 81% after depuration, a 1,800-fold increase in sequence reads specific to *Vibrio*. This dramatic increase suggests that the genus *Vibrio* is not as effectively removed as other bacteria, and as a consequence make up a larger fraction of the total microbiome in the depurated oyster. *Vibrio* species that remain after depuration may be a part of the oyster’s permanent flora.

Five bacterial families were no longer detectable following depuration: *Bacillaceae*, *Staphylococcaceae*, *Enterococcaceae*, *Rhodobacteraceae*, and *Methyllococcaceae* (Table 4.4). The largest decrease in percentage of sequences was from the bacterial family *Rhodobacteraceae* whose members made up 2.1% before. All other bacterial families found in the oyster prior to
depuration each contained less than 0.6% of sequence reads and were not detected after depuration suggesting these organisms were part of the transient flora of C. virginica.

Discussion

In this study, SEC was used to obtain information on bacterial diversity by capturing and retrieving 16S rRNA genes with the intent of establishing whether SEC could serve as a method for obtaining taxonomic information from environmental samples. Three distinct, but interrelated, marine environments were investigated: the cultured oyster C. virginica, seawater above the oyster bed, and sediment beneath the oysters. The oyster hatchery is located in an estuary near Grand Isle, LA and is part of the Gulf of Mexico. Marine environments, particularly seawater and oysters, are a challenge when using metagenomic sequencing as a means of defining the bacterial component of these environments. Marine environments can include large numbers of eukaryotes that disproportionally contribute to DNA isolated from these environments, potentially masking the contributions from bacterial DNA.

Prokaryotes found in marine environments outnumber eukaryotes with concentrations of \(10^4\)-\(10^7\) prokaryotic cells/mL (Whitman, Coleman et al. 1998) versus \(10^2\)-\(10^4\) eukaryotic cells/mL (Diez, Pedros-Alio et al. 2001). In general, genome sizes of marine eukaryotes are one to two orders of magnitude larger than the bacteria found in these environments. For example, dinoflagellates and single-cell algae have genome sizes between 3-245 x \(10^6\) kbp and 13-200 x \(10^6\) bp (Hou and Lin 2009), respectively, compared to an average bacterial genome of 4 x \(10^6\) bp. Thus, the majority of the DNA isolated from seawater is likely eukaryotic. SEC was effective in capturing 16S rRNA genes from filtered seawater surrounding oyster cages. Seventeen bacterial families were identified from this seawater sample (Table 4.1) have also been previously reported elsewhere in the Gulf of Mexico: Flavobacteriaceae, Alteromonadaceae,
Microbacteriaceae, Enterobacteriaceae, Cryomorphaceae, Saprospiraceae, and Chamaesiphonaceae (Felder 2009, Newton, Huse et al. 2013).

Bacterial diversity identified using SEC captured 16S rRNA genes from Grand Isle seawater was similar to diversity defined in seawater collected from St. George Island, FL, a site also located in the Gulf of Mexico at similar latitude (Figure 4.1). The Grand Isle and St. George Island samples (Table 4.1) have seven bacterial families in common. More bacterial families were identified with SEC (17 families) in the Grand Isle samples as were found on St. George Island using pyrotags (Newton, Huse et al. 2013). While this difference may be due to compositional variation in the communities at these sites, SEC coupled with the EMIRGE program provided longer 16S rRNA gene sequences compared to the pyrotags from Newton et al. 2013.

SEC was also successful in retrieving 16S rRNA genes from marine sediment, identifying more bacterial diversity in sediment than seawater, 31 versus 17 bacterial families respectively. Many assembled sequences from SEC-Sediment belonged to facultative and strict anaerobes. The dominant families, Streptococcaceae and Desulfobacteraceae, contained the genera Streptococcus and Desulfoluna. Streptococcus has been isolated as potential fecal contaminants from both fresh and marine sediments (Erkenbrecher 1981, Davies, Long et al. 1995) and results from this study suggest fecal contamination at this oyster hatchery near Grand Isle, LA, but since the Streptococcaceae are ubiquitously distributed, further investigation is needed to determine the source of these bacteria. The presence of the genus Cetobacterium (a member of Fusobacteriaceae) also indicates a possible fecal contaminant (Finegold, Vaisanen et al. 2003). Fusobacteriaceae have also been discovered in sediment from a tidal flat in the Wadden Sea, Germany, a mangrove swamp in China, and coastal sediment from Tokyo Bay.
(Kopke, Wilms et al. 2005, Liang 2007, Matsui 2013), but they have never before been reported in coastal sediment from the Gulf of Mexico. SEC also identified several bacterial families with members that have sulfate-reducing (SR) capabilities, Desulfomicrobiaceae and Desulfobulbaceae. Several SR bacterial families have been found in marine sediments in the Gulf of Mexico. These studies were collected at St. George Island, FL (Mills, Hunter et al. 2008), and at a methane seep (Lloyd, Lapham et al. 2006). In addition, the genus Nitrospina, which includes nitrite oxidizers, was identified; no evidence of members from Nitrospinaceae has been found in coastal sediment from the Gulf of Mexico, previously.

Comparisons between marine sediment near St. George Island, FL and the sediment sample here revealed considerable differences in bacterial diversity. Of the 38 families identified at the two sites, only five were common to both (Table 4.2). Thus, while the bacterial diversity in seawater observed in Grand Isle and St. George Island were similar, differences were much more pronounced in marine sediment. The two sampling sites were not identical, and the sediment may have provided a more location-specific signature than did the seawater. Grand Isle sediment is from an estuary, whereas the St. George Island sample was taken under open water. The sites presumably had different geochemistries, and it is assumed that those differences are reflected in the microbial composition at each site.

In this study, genera were identified within marine sediment that may be capable of methane oxidation and methylotrophy, suggesting that methanogens are present in the sediment. Felder et al. (2009) predicted at least 50 species of archaea likely to exist in the Gulf of Mexico and members from Methanosarcinales and Methanomicrobiales have been identified in sediments near a methane seep in the Gulf of Mexico (Lloyd, Lapham et al. 2006). No archaeal 16S rRNA gene sequences were reassembled from any of the datasets. Analysis of the
biotinylated Uni1390R primer against the SILVA database showed this primer was complementary to only 5.3% of archaeal sequences in the database (Klindworth, Pruesse et al. 2013). Improvements that would increase our chance of detecting archaea could be made by modifying the biotinylated capture primer sequence to more effectively retrieve both archaeal and bacterial 16S rRNA genes. Klindworth et al. 2013 provides analysis of the primer S-D-Bact-0785-a-A-18 that matches to both Archaea and Bacteria only at 96.8 and 96.5% when allowing for a single mismatch and could be used as an alternative biotinylated capture primer specific for prokaryotes.

SEC was tested for its ability to capture 16S rRNA genes within *C. virginica* in an effort to describe the bacterial diversity within this eukaryote. Despite the overabundance of oyster DNA, assembled sequences were bacteria from two phyla, Proteobacteria and Firmicutes (Figure 4.3). A majority of the assembled sequences from the oyster were identified as the genera *Escherichia, Streptococcus*, and *Enterococcus* and could be the consequence of fecal contamination of water near the oyster hatchery. Fecal contaminants have been previously reported in *C. virginica* using culture-dependent methods (Slanetz and Bartley 1964, Hood, Ness et al. 1983, Hoi, Larsen et al. 1998, Pfeffer, Hite et al. 2003), with some methods focusing specifically on enumeration of fecal streptococci (Slanetz and Bartley 1964). SEC could potentially be used as a culture-independent method of detecting fecal coliforms in oysters.

Other bacterial genera identified within SEC-Oyster were *Bacillus* and *Staphylococcus* and have been reported in *C. gigas*, a related species to *C. virginica* (Geiger, Ward et al. 1926, Geiger and Crowley 1937, Colwell and Liston 1960). Other 16S rRNA genes assembled belonged to the genus *Rhodobacter* which have been previously reported as symbionts of the marine sponge, *Halichondria panacea* (Althoff, Schuett et al. 1998) and the American lobster
(Lavallee, Hammell et al. 2001), respectively; their presence in *C. virginica* had not been previously reported.

The microbiome of *C. virginica* identified through SEC was compared to a previous study that used pyrotags to determine the prokaryotic diversity of the stomach and gut of *C. virginica* harvested from a location in the same estuary as Grand Isle oyster hatchery (King, Judd et al. 2012). There were differences in the total number of bacterial phyla identified using pyrotags versus SEC (Figure 4.3). The majority of the phyla reported by King, Judd et al. (2012) were not detected in the whole oyster homogenates used in this study. Bacteria from the phyla Verrucomicrobia, Planctomycetes, and Actinobacteria, have been reported in samples taken from an oyster shell dump, suggesting their associations with oyster shells (Math, Islam et al. 2010) and were also found in King, Judd et al. (2012). From this study, sequences for members of Verrucomicrobia and Planctomycetes were found in Grand Isle sediment, but were not identified in the oyster. Planctomycetes have also been previously found in biofilms near oyster reefs (Nocker, Lepo et al. 2004). As will be discussed below, much of the flora associated with the oyster appears to be transient. As filter feeders, their gut flora will reflect what is in the water before harvest. In other words, as the composition of the microbial flora ingested by the oyster changes, it is expected that the oyster’s transient microbiome will change accordingly. If these bacteria were not present in the oyster gut analyzed by King, Judd et al. (2012), the differences in microbial diversity are easily explained. If we assume that these bacteria were present at the time of collection for King, Judd et al. study, we must conclude that their relative abundance was low enough to escape detection. As with the possible explanations offered for differences between the seawater and sediment samples found at different locations, there is no unequivocal method to account for the differences without additional study.
It is noted that the methodology used in the previous characterization of the oyster gut (King, Judd et al. 2012) was not identical to that used here. The two studies differed in post-harvest temperature; in this study, oysters were held at room temperature whereas in the study by King, Judd et al. (2012), oysters were held at 4 °C. A study on bacteria within C. gigas found that different post-harvest temperatures could dramatically affect bacterial diversity observed within the oyster (Fernandez-Piquer, Bowman et al. 2012). Perhaps this methodological difference affected the diversity observed in the oyster homogenates described in our study.

A large amount of research investigating bacteria that inhabit C. virginica has involved the detection of Vibrio species through cultivation or quantitative PCR (Wright, Miceli et al. 1993, Wright, Hill et al. 1996, Cerda-Cuellar, Jofre et al. 2000, Campbell and Wright 2003, Chae, Cheney et al. 2009). Vibrio species have been identified in the seawater and sediment surrounding C. virginica (Wright, Hill et al. 1996, Pfeffer, Hite et al. 2003, Panicker, Myers et al. 2004, Zimmerman, DePaola et al. 2007, Johnson, Flowers et al. 2010). Of the 283 assembled 16S rRNA gene sequences from the SEC-Oyster dataset, only two were identified as Vibrio and represented 0.53% of the sequence reads. Enumeration of Vibrio in C. virginica has been shown to vary based on salinity and water temperature (Wright, Hill et al. 1996, Johnson, Flowers et al. 2010). Our results found Vibrio present in the oyster, but undetectable in the seawater and sediment surrounding the oysters, but they may have been below our limits of detection. This study is the first to detect Vibrio in C. virginica without amplification using high-throughput sequencing.

This work represents the first attempt to correlate the microbiome of C. virginica with the surrounding seawater and sediment. A single family (Enterobacteriaceae) was shared between all three environments. The assembled sequences belonged to the genus Escherichia. Since
*Escherichia* species are normally found in the mammalian gut, this finding indicates fecal contamination of the environment where the oysters were harvested. A recent study investigating microbial diversity of a public beach in Grand Isle found the site to be dominated by *Escherichia* (Engel and Gupta 2014). *E. coli* has also been shown to be actively taken up and retained by *C. virginica* from seawater (Murphree and Tamplin 1995). Of the three environments, sediment and seawater shared the most bacterial diversity. Five families (*Alcanivoracaceae, Cryomorphaceae, Alteromonadaceae, Chromatiaceae* and *Saprospiraceae*) were detected in these environments and their appearance is likely due to the mixing of water and sediment by water currents.

The bacteria found in the sediment most closely resembled those found in the oyster (Figure 4.4). During growth stages, *C. virginica* naturally form oyster beds on the sediment. When temperatures are warm they release their sperm and eggs for fertilization (Banks 2007). After fertilization, the larvae exhibit anti-phototaxis and attach to a solid surface where nutrients are present (Banks 2007). It is reasonable to assume that the larvae will remain close to the sediment to avoid light and filter feed the seawater in close proximity to the sediment. It may be that during the juvenile stages, oysters filter the seawater in close proximity with the sea floor and develop residential flora that includes some of the species found in the sediment. It has also been shown that chemical inducers produced by bacteria like *Vibrio* and *Escherichia* increased settlement behavior of the Pacific oyster, *C. gigas* larvae (Fitt, Labare et al. 1989). *Escherichia* was also identified through SEC in the sediment while *Vibrio* was only found in the oyster. There is a possibility that larvae will develop residential flora more similar to microorganisms found in the sediment that remain permanent members through adulthood.
SEC was also used to track changes in the microbiome of *C. virginica* following depuration treatment. Studies of the changes in microbial diversity following depuration treatment in oysters focused have on coliforms and pathogenic *Vibrio* species (Son and Fleet 1980, Froelich, Ringwood et al. 2010, Froelich and Oliver 2013). To our knowledge, no study has reported changes in other bacteria present in the microbiome. SEC provided a way to monitor how the microbiome of *C. virginica* changes following depuration treatment. Overall, there was a decrease in the number of bacterial families identified from thirteen to three following depuration treatment (Figure 4.5). This loss presumably occurs because transient bacteria are flushed out of the digestive tract. Decreases in *Escherichia*- and *Enterococcus*-associated colony forming units have been reported in *C. virginica* after depuration, but the complete removal of these organisms was not achieved (Love, Lovelace et al. 2010). In addition, fecal *Escherichia* and *Streptococcus* have also been found in *C. virginica* tissues (Hood, Ness et al. 1983, Hoi, Larsen et al. 1998) and *Escherichia* has been reported to remain in oysters at much lower levels following depuration treatment (Love, Lovelace et al. 2010).

The most notable difference between the microbiome of *C. virginica* before and after depuration treatment was in the assembled sequences for the genus *Vibrio*. That fraction increased from 0.53% to 81% following depuration treatment (Table 4.4). While depuration allows the oyster to purge itself of bacteria, reducing the number and types of bacteria present, our results indicate that *Vibrio* species remained associated with oysters. A recent report argues that *C. virginica* can eliminate exogenous added laboratory strains of *V. vulnificus* following depuration, but resident *V. vulnificus* remains in the oyster (Froelich and Oliver 2013). Significant numbers of these resident *V. vulnificus* are unaffected by depuration, and their numbers increase during treatment (Froelich and Oliver 2013). This scenario may explain what
was observed in this study. In addition, there is evidence that *V. cholerae* also persists in oysters after depuration (Murphree and Tamplin 1995). It is possible that depuration results in the loss of transient bacteria and as a result, *Vibrio* species were able to grow without competition, dominating the oyster’s microbiome. These results demonstrate that many of the bacteria inhabiting *C. virginica* could be transient and some like *Vibrio* and *Escherichia* may be resistant to removal post-depuration.

These results demonstrate the effectiveness of SEC in capturing bacterial 16S rRNA genes from *C. virginica* even after depuration treatment. When the detection of a specific bacterium is warranted, SEC could be used to target specific pathogenic bacteria inhabiting the oysters. Other pathogenic genera found in oysters include *Salmonella* (Brands, Inman et al. 2005), *Clostridium* (Muniaín-Mujika, Calvo et al. 2003), and *Campylobacter* (Teunis, Havelaar et al. 1997). A biotinylated primer could be designed that is specific to the bacterium of interest with the SEC protocol being implemented.

These results further demonstrate that SEC can be used to target and retrieve 16S rRNA genes from environmental samples. SEC coupled with EMIRGE permits nearly full-length 16S rRNA gene sequences. The utility of SEC in retrieving 16S rRNA genes in a eukaryote can further our understanding of microbiomes in complex biological systems without the need for extensive sequencing as would be the case with metagenomic sequencing projects.
CHAPTER 5.
MICROBIAL DIVERSITY OF BASAL ICE FROM TAYLOR GLACIER THROUGH SEC CAPTURE OF THE 16S rRNA GENE

Introduction

The term ‘rare biosphere’ has been applied to microorganisms believed to be a part of the Earth’s microbiome. These microbes are potentially vital members of the community, but because they are present in such low abundance, current methods fail to detect them (Sogin, Morrison et al. 2006). Like the rare biosphere, low biomass environments pose problems when attempting to analyze microbial diversity. Low biomass means low recoveries of DNA from the metagenome present. The resulting low DNA yield can affect the PCR amplification of low abundant templates in library preparations prior to high-throughput sequencing (Kennedy, Hall et al. 2014). DNA concentrations sufficiently large enough to support high-throughput sequencing must be available when characterizing these environments. To illustrate this point, consider polar environments like those found in subglacial lakes and glaciers. Some polar environments contain as few as $10^2$ cells/gram of ice (Doyle, Montross et al. 2013). Conventional techniques for isolating DNA yields concentrations that do not meet the requirements of high-throughput sequencing platforms when performing metagenomic sequencing. DNA extraction methods can also impact DNA yield (La-Duc 2009) resulting in lower than expected yields that hinder high-throughput sequencing. To avoid this problem, PCR or whole genome amplification is used to increase DNA concentrations to a level sufficient for analysis. The standard input DNA amounts for high-throughput sequencing makes PCR amplification a viable option to analyze low biomass environments.

PCR amplification has been used for the characterization of microbial communities. Currently, an experimental paradigm used involves generating iTags that are sequenced. iTags
are PCR amplified regions of the 16S rRNA gene sequenced on Illumina platforms (Degnan and Ochman 2012). While different variable regions may be targeted for amplification, a well-documented protocol for the V4 region of the 16S rRNA gene (Caporaso, Lauber et al. 2011) has been widely accepted for analysis of environmental samples, being implemented in the Earth Microbiome Project (http://www.earthmicrobiome.org).

In this study, we explore the capacity of SEC to facilitate characterization of bacterial diversity of a low biomass environment. We hypothesize that combining the utility of SEC in capturing low abundant sequences (Chapter 3) with the successful implementation of SEC on environmental samples (Chapter 4) will provide a method that is superior to that applied in these environments. The microbial community chosen for this study is located in the basal ice of the Taylor Glacier in Antarctica. This ice contains microbial cells at concentrations between $10^2$-$10^4$ cells gram$^{-1}$ of ice and exhibits metabolic activity at -15°C (Doyle, Montross et al. 2013). A previous study generated a clone library from this basal ice that included 43 clones representing five genera: Bacillus (21), Cohnella (6), Paenisporosarcina (6), Acinetobacter (9), and Psychrobacter (1) (Doyle, Montross et al. 2013). It is anticipated that SEC, because of its ability to capture the entire 16S rRNA gene, will provide a broader characterization of the bacterial diversity in this environment relative to V4 tags.

**Results**

**SEC Retrieved 16S rDNA from a Low Biomass Environment**

The microbial diversity of low biomass environments, including glaciers, are routinely characterized using clone libraries generated from PCR amplified 16S rRNA genes (Christner, Mosley-Thompson et al. 2003, Skidmore, Anderson et al. 2005, Mikucki and Priscu 2007, Doyle, Montross et al. 2013). More recently, PCR amplicon sequencing with pyrotags has also
been used to determine the microbial diversity for glacier environments (Simon, Wiezer et al. 2009, Schutte, Abdo et al. 2010, Pearce, Hodgson et al. 2013, Shtarkman, Kocer et al. 2013). In this study, SEC was used to target and retrieve the 16S rRNA genes from the metagenome found in basal ice from Taylor Glacier in Antarctica. Genomic DNA (21.4 pg/μL) extracted from the basal ice was provided by Shawn Doyle of the Christner Research Group (Department of Biological Sciences, Louisiana State University). This DNA was obtained from microorganisms collected following filtration of melted basal ice (Doyle, Montross et al. 2013). SEC was performed on 105 pg of genomic DNA at a final concentration of 2.1 pg/μL. The captured biotinylated product was then sequenced, avoiding PCR amplification prior to library preparations. The Taylor Glacier dataset (SEC-TG) was trimmed using CLC Genomics Workbench and resulted in 76,846387 reads. The dataset was subsampled three times using a custom script (Miller 2013) and analyzed using EMIRGE. EMIRGE recreated 600 ± 120 16S rRNA gene sequences with an average length of 1465 ± 7 base pairs.

Three bacterial phyla were classified, Firmicutes, the most prevalent, contained 45% of the sequence reads from assembled sequences, followed by Actinobacteria (40%), and Proteobacteria (15%) (Figure 5.1, SEC-TG). Previous work by Doyle et al. (2013) found culturable bacteria belonging to Firmicutes from the same basal ice sample used in this study. In addition, a small clone library derived from genomic DNA isolated from the same sample revealed sequences from Proteobacteria and Firmicutes (Doyle, Montross et al. 2013).
Eleven bacterial families were identified. Seven families dominated, making up 92% of the total sequence reads, *Bacillaceae* (34.72 ± 6.93%), *Actinomycetaceae* (22.82 ± 12.55%), *Pseudomonadaceae* (11.70 ± 6.35%), *Micrococcaceae* (7.86 ± 3.28%), *Paenibacillaceae* (6.75 ± 2.95%), *Propionibacteriaceae* (4.17 ± 3.79%), and *Pseudonocardiaceae* (4.58 ± 4.11%) (Table 5.1). The remaining four families contained sequence reads that ranged between 0.12-2.0% of the total and accounted for the remaining 4.6% of the total sequence reads.

Through the use of EMIRGE, full-length 16S rRNA genes were obtained facilitating genus level identification. Thirteen genera were identified. The most abundant were *Bacillus*, *Actinomyces*, *Pseudomonas*, *Arthrobacter*, *Cohnella*, *Kocuria*, *Propionibacterium*, and *Pseudonocardia*, accounted for 93% of the sequence reads (Table 5.1). Four of the genera identified from SEC-TG were also found in the clone library previously generated from this basal ice (Doyle, Montross et al. 2013); those genera included *Bacillus*, *Cohnella*, *Acinetobacter*, and *Paenisorosarcina* which makes up 44% of the sequence reads (Table 5.1).
Table 5.1 Phylogenetic composition of SEC-TG analyzed with EMIRGE.

<table>
<thead>
<tr>
<th>Phylum</th>
<th>Class</th>
<th>Family</th>
<th>Genus</th>
<th>SEC-TG</th>
</tr>
</thead>
<tbody>
<tr>
<td>Actinobacteria</td>
<td>Actinobacteria</td>
<td>Actinomycetaceae</td>
<td>Actinomyces</td>
<td>22.82 ± 12.55</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Intrasporangiaceae</td>
<td>Ornithinimicrobium</td>
<td>0.13 ± 0.12</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Micrococcaceae</td>
<td>Arthrobacter</td>
<td>4.90 ± 0.57</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Kocuria</td>
<td>2.96 ± 2.71</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Propionibacteriaceae</td>
<td>Propionibacterium</td>
<td>4.17 ± 3.79</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Pseudonocardiaceae</td>
<td>Pseudonocardia</td>
<td>4.58 ± 4.11</td>
</tr>
<tr>
<td>Firmicutes</td>
<td>Bacilli</td>
<td>Bacillaceae</td>
<td>Bacillus</td>
<td>34.72 ± 6.93</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Paenibacillaceae</td>
<td>Cohnella</td>
<td>6.08 ± 2.42</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Planococcaceae</td>
<td>Paenisporosarcina</td>
<td>0.67 ± 0.53</td>
</tr>
<tr>
<td></td>
<td>Clostridia</td>
<td>Clostridiaceae</td>
<td>Clostridium</td>
<td>0.32 ± 0.30</td>
</tr>
<tr>
<td>Proteobacteria</td>
<td>γ-Proteobacteria</td>
<td>Enterobacteriaceae</td>
<td>Escherichia</td>
<td>1.46 ± 0.33</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Moraxellaceae</td>
<td>Acinetobacter</td>
<td>2.00 ± 1.62</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Pseudomonadaceae</td>
<td>Pseudomonas</td>
<td>11.70 ± 6.35</td>
</tr>
</tbody>
</table>

Note: Values are shown as percentages calculated by EMIRGE and are based on the number of sequence reads attributing to assembled sequences.

**Microbial Characterization of V4 iTags from Basal Ice of Taylor Glacier**

iTags provide an alternative method to determine the microbial diversity. In this study, we generated and sequenced V4 iTags from organisms found in the basal ice from Taylor Glacier. The V4 iTags described here were generated as a part of a collaboration with the Christner Research Group (Department of Biological Sciences, Louisiana State University) using the protocol outlined by Caporaso et al. (2011) and were sequenced using the Illumina MiSeq platform. Generation of the V4 tags was performed by Shawn Doyle. The V4 iTags were analyzed using the program Mothur (Schloss, Westcott et al. 2009) version 1.33.3 ([http://www.mothur.org/](http://www.mothur.org/)) to remove reads with ambiguous bases, homopolymers of >8 bases, sequence read lengths of <143 bases, and which had an average quality score of <35. UCHIME (Edgar, Haas et al. 2011) was used to detect and remove chimeras. Post-processing, 58,735 sequence reads were available and provided to me for analysis. The V4 iTags were analyzed using the software package QIIME (Caporaso, Kuczynski et al. 2010) to determine the bacterial diversity. Phylogenetic composition was determined using the de novo OTU picking workflow.
within QIIME. Four percent of the OTUs were not classified at the phylum level (Figure 5.1). A total of sixteen bacterial phyla were identified; the majority of OTUs were Firmicutes (44.7%), followed by Proteobacteria (43.6%), Bacteroidetes (5.9%), Actinobacteria (1.1%), Deinococcus-Thermus (0.16%) and Cyanobacteria (0.50%). The remaining ten phyla, Acidobacteria, Armatimonadida, Caldiserica, Chloroflexi, Gemmatimonadetes, Verrucomicrobia, and candidate phyla NC10, OP8, OP9, and WPS-2 contained less than 0.25% of the total classified OTUs at the phylum level (Figure 5.1).

The bacterial composition at the family level resulted in a majority of the OTUs being unclassified (64% of the total OTUs) and are not reported in Table 5.2. The OTUs that were classified at the family level identified forty-three bacterial families from eleven phyla (Table 5.2). The most dominant family was Moraxellaceae at 87% of the classified OTUs followed by Bacillaceae (2.7%), Comamonadaceae (2.6%), Flexibacteraceae (1.8%), Paenibacillaceae (1.7%), Methylophilaceae (0.8%), Acetobacteraceae (0.5%), Deinococcaceae (0.4%), Chitinophagaceae (0.4%), and Chamaesiphonaceae (0.3%) (Table 5.2). The remaining thirty-two families contained OTUs that ranged from 0.01 to 0.16% and additively was less than 1.8% of the total classified OTUs. Clones from the same basal ice sample used for the V4 iTags also found bacterial families Bacillaceae, Moraxellaceae, Paenibacillaceae (Doyle, Montross et al. 2013).

Table 5.2 Phylogenetic composition of V4 tags analyzed with QIIME.

<table>
<thead>
<tr>
<th>Phylum</th>
<th>Class</th>
<th>Family</th>
<th>V4 iTags</th>
</tr>
</thead>
<tbody>
<tr>
<td>Acidobacteria</td>
<td>Acidobacteria</td>
<td>RB40</td>
<td>0.01</td>
</tr>
<tr>
<td></td>
<td></td>
<td>ACK-M1</td>
<td>0.07</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Intrasporangiaceae</td>
<td>0.01</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Microbacteriaceae</td>
<td>0.31</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Micrococcaceae</td>
<td>0.11</td>
</tr>
<tr>
<td>Actinobacteria</td>
<td>Actinobacteria</td>
<td>Sporichthyaceae</td>
<td>0.03</td>
</tr>
<tr>
<td>Phylum</td>
<td>Class</td>
<td>Family</td>
<td>V4 iTags</td>
</tr>
<tr>
<td>------------------</td>
<td>------------------------</td>
<td>---------------------</td>
<td>----------</td>
</tr>
<tr>
<td>Armatimonadetes</td>
<td>Armatimonadaria</td>
<td>WD294</td>
<td>0.06</td>
</tr>
<tr>
<td>Bacteroidetes</td>
<td>Flavobacteria</td>
<td>Flavobacteriaceae</td>
<td>0.04</td>
</tr>
<tr>
<td></td>
<td>Chitinophagaceae</td>
<td>0.36</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Flexibacteriaceae</td>
<td>1.75</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Sphingobacteriaceae</td>
<td>0.04</td>
<td></td>
</tr>
<tr>
<td>Chloroflexi</td>
<td>Anaerolineae</td>
<td>Anaerolinaceae</td>
<td>0.15</td>
</tr>
<tr>
<td>Cyanobacteria</td>
<td>Nostocophycideae</td>
<td>Nostocaceae</td>
<td>0.01</td>
</tr>
<tr>
<td></td>
<td>Synechococcophycideae</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Pseudanabaenaceae</td>
<td>0.1</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Chamaesiphonaceae</td>
<td>0.3</td>
</tr>
<tr>
<td>Firmicutes</td>
<td>Bacilli</td>
<td>Bacillaceae</td>
<td>2.66</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Paenibacillaceae</td>
<td>1.68</td>
</tr>
<tr>
<td></td>
<td>Clostridia</td>
<td>Clostridiaceae</td>
<td>0.02</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Peptococcaceae</td>
<td>0.02</td>
</tr>
<tr>
<td>Gemmatimonadetes</td>
<td>Gemmatimonadetes</td>
<td>A1-B1</td>
<td>0.04</td>
</tr>
<tr>
<td>Proteobacteria</td>
<td>α-Proteobacteria</td>
<td>Caulobacteriaceae</td>
<td>0.03</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Bradyrhizobiaceae</td>
<td>0.01</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Rhodobacteriaceae</td>
<td>0.03</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Acetobacteriaceae</td>
<td>0.5</td>
</tr>
<tr>
<td></td>
<td>β-Proteobacteria</td>
<td>Alocigenaceae</td>
<td>0.01</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Comamonadaceae</td>
<td>2.64</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Oxalobacteriaceae</td>
<td>0.16</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Hydrogenophilaceae</td>
<td>0.03</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Methylphilaceae</td>
<td>0.84</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Rhodocyclaceae</td>
<td>0.14</td>
</tr>
<tr>
<td></td>
<td>δ-Proteobacteria</td>
<td>Bacteriovoracaceae</td>
<td>0.01</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Desulfobulbaceae</td>
<td>0.07</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Haliangiaceae</td>
<td>0.01</td>
</tr>
<tr>
<td>Proteobacteria</td>
<td>γ-Proteobacteria</td>
<td>Alteromonadaceae</td>
<td>0.01</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Enterobacteriaceae</td>
<td>0.17</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Crenotrichaceae</td>
<td>0.14</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Moraxellaceae</td>
<td>86.51</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Pseudomonadaceae</td>
<td>0.2</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Piscirickettsiaceae</td>
<td>0.11</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Xanthomonadaceae</td>
<td>0.15</td>
</tr>
</tbody>
</table>
Comparison Between V4 Tags and SEC-TG

Figure 5.2 shows a Venn diagram displaying the bacterial families identified through V4 tags and SEC-TG. Eight bacterial families were identified by both methods: Bacillaceae, Paenibacillaceae, Clostridiaceae, Intrasporangiaceae, Enterobacteriaceae, Moraxellaceae, Pseudomonadaceae, and Micrococcaceae. These eight bacterial families contained 92% of the sequence reads in SEC-TG and 93% of the classified OTUs for the V4 iTags, suggesting that these bacterial families could be the dominant members in the community.

Three bacterial families were unique to the SEC-TG dataset and made up 32% of the sequence reads with the most abundant family of this group being Actinomycetaceae that contained 23% of the sequence reads (Table 5.1). Thirty-two distinct bacterial families were identified in V4 iTags (Figure 5.2) and accounted for 8.7% of the classified OTUs, indicating a relatively low number of sequence reads attributing to their classification.
Discussion

Characterization of microbial community diversity of a low biomass environment through high-throughput sequencing offers a window into these biospheres; the low yields of DNA obtained from the subsurface environments can adversely affect the sequencing platforms which require substantial DNA for normal operation. PCR amplification of variable regions of the 16S rRNA gene has been used to achieve DNA concentrations sufficient for high-throughput sequencing (Pearce, Hodgson et al. 2013). Although PCR amplicon sequencing promises to reveal the “rare biosphere” (Sogin, Morrison et al. 2006), PCR biases and sequencing errors can result in over- or underestimates of species diversity (Kunin, Engelbrektsen et al. 2010, Jumpstart Consortium Human Microbiome Project Data Generation Working 2012).

In this study, we assessed whether SEC could retrieve 16S rRNA genes from a low biomass environment without PCR amplification. Using the microbial community found in basal ice taken from Taylor Glacier in Antarctica, we demonstrated that SEC could recover these sequences. Despite the low cell concentration, SEC identified members of three bacterial phyla, Firmicutes, Actinobacteria, and Proteobacteria (Figure 5.1). Several studies investigating the microbial diversity of glacier environments using pyrotags have reported greater bacterial diversity (Pearce, Hodgson et al. 2013, Shtarkman, Kocer et al. 2013).

SEC combined with EMIRGE allowed for recreation full-length 16S rRNA genes with an average length of 1465 ± 7.6 bases. Twelve bacterial families and thirteen genera were identified (Table 5.1). The dominant families classified from this study, *Bacillaceae, Micrococcaceae, Paenibacillaceae, Moraxellaceae, Micrococcaceae*, and *Pseudomonadaceae*, have also been reported in ice originating from China, Bolivia, Antarctica, and the Greenland Ice Sheet Project (Christner, Mosley-Thompson et al. 2000, Miteva, Sheridan et al. 2004). Members
of *Bacillaceae, Pseudomonadaceae, and Clostridiaceae* were found using pyrosequencing of DNA from accretion ice from Lake Vostok (Shtarkman, Kocer et al. 2013) suggesting these bacterial families are widely distributed in cold environments. Of the thirteen genera identified from SEC-TG, *Bacillus, Arthrobacter, and Acinetobacter* contained 42% of the sequence reads from assembled sequences and were also reported from accretion ice from Lake Vostok (Christner, Mosley-Thompson et al. 2000) providing evidence of their presence in cold environments. *Pseudomonas, Bacillus, and Clostridium* identified from SEC-TG have also been reported from ice cores from the Greenland Ice Sheet Project, sediment from Hodgson Lake, and accretion ice from Lake Vostok (Christner, Mosley-Thompson et al. 2001, Sheridan, Miteva et al. 2003, Pearce, Hodgson et al. 2013, Shtarkman, Kocer et al. 2013).

V4 iTags were generated on the same DNA sample used for SEC by Shawn Doyle of the Christner Research Group. They were sequenced and analyzed using QIIME. Sixteen phyla were identified (Figure 5.1). Similar results have been reported at the phylum level from studies investigating the microbial diversity from several other glaciers. A clone library (133 clones) derived from basal ice from John Evans Glacier, Canada reported Proteobacteria, Bacteroidetes, Actinobacteria, Verrucomicrobia, and Acidobacteria (Skidmore, Anderson et al. 2005). Soil samples from the foreland of Midre Loven glacier, West Spitsbergen, Norway used pyrotags and reported the following bacterial phyla Acidobacteria, Actinobacteria, Cyanobacteria, Chloroflexi, Deinococcus-Thermus, Bacteroidetes, Firmicutes, Gemmatimonadetes, Verrucomicrobia, and Proteobacteria (Schutte, Abdo et al. 2010), all of which were also found in V4 iTags. In addition, Cyanobacteria, Proteobacteria, Bacteroidetes, Actinobacteria, Gemmatimonadetes, Acidobacteria, Deinococcus-Thermus, and Firmicutes were reported in association with the Schneeferner glacier in Germany (Simon, Wiezer et al. 2009). Although these studies were
conducted at different glaciers around the world from different sampling sites including glacial ice, basal ice, and in the foreland of the glacier, they all exhibit evidence of similar profiles among the bacterial phyla reported following analysis of the V4 iTags from basal ice from Taylor Glacier.

Only 36% of OTUs derived from V4 iTags could be classified into families with 42 bacterial families identified (Table 5.2). These 42 families have never been identified in a single sample before, but smaller groups of families were identified together in other glaciers. Ice collected at Guliya, China, Sajama, Bolivia, and Taylor Dome, Antarctica reported groupings of *Moraxellaceae, Bacillaceae, Flavobacteriaceae, Microbacteriaceae, Micrococcaceae,* and *Paenibacillaceae* (Christner, Mosley-Thompson et al. 2000), and members of *Paenibacillaceae, Microbacteriaceae, Pseudomonadaceae, Xanthomonadaceae, Micrococcaceae,* and *Moraxellaceae* have been cultured from ice cores from the Greenland Ice Sheet Project (Miteva, Sheridan et al. 2004). Members of *Paenibacillaceae, Comamonadaceae, Pseudomonadaceae,* and *Sphingobacteriaceae* were also isolated from an ice core from Lake Vostok (Christner, Mosley-Thompson et al. 2001). Several families that were identified offered clues on the metabolic diversity of the bacterial families found in the basal ice of Taylor Glacier. The presence of OTUs specific for *Anaerolinaceae* suggested the presence of green nonsulfur bacteria. OTUs specific for *Methylophilaceae* and *Desulfobulbaceae* provide evidence for bacteria capable of methylotrophy and sulfate-reducing capabilities (Castro, Williams et al. 2000, Lapidus, Clum et al. 2011).

Chapter 3 demonstrated SEC was effective in capturing not only 16S rRNA gene sequences from an artificial bacterial mock community, but also more successfully predicted relative abundance of the members within an order of magnitude compared to PCR amplicons of
the V3-V5 region of the 16S rRNA gene. A comparison for the effectiveness of SEC in predicting the relative abundance with the V4 iTags of the basal ice was not possible as QPCR analysis of any of the members was not performed. However, both methods contain members of the community present at abundances of greater than 20% (Table 5.1, Table 5.2). Furthermore, both SEC and V4 iTags shared eight bacterial families that accounted for 92 and 93% of the assembled sequences and OTUs classified, respectively (Table 5.1, Table 5.2, Figure 5.2). This suggests these bacterial families are dominant members of the community.

A previous study investigating the microbial diversity in basal ice from Taylor Glacier cultured members of the genus *Paenisporosarcina* from the family *Plancococcaceae* (Doyle, Montross et al. 2013). Neither this genus or family was identified in V4 iTags generated in this study, but evidence for this group’s existence was found using SEC and EMIRGE. In other words, SEC was able to retrieve a known member of the basal ice that V4 iTags failed to detect. Doyle et al 2013 also reported *Psychrobacter* in the same basal ice but this genus was not identified in either SEC-TG or V4 iTags, suggesting that neither method captures all the bacterial diversity exhibited in this low biomass environmental sample.

When combined SEC/EMIGE and V4 iTags classified 47 bacterial families. Thirty-eight of those families were identified by only one of the two methods (Figure 5.2). The V4 iTags classified a total of 43 families, 35 of which were solely classified by V4 iTags (Figure 5.2). At the family level, eighty-seven percent of the classified OTUs were identified as *Moraxellaceae* while the remaining 13% were identified across 42 bacterial families suggesting a bias in amplification of the family *Moraxellaceae*. PCR analysis of 16S rRNA genes can result in differential amplification (Suzuki and Giovannoni 1996, Polz and Cavanaugh 1998) and perhaps the V4 iTags could have contained an over amplification of members of *Moraxellaceae* that was
not observed in SEC-TG. Chimeras were detected in the V4 iTag dataset (Shawn Doyle, personal communication) and could indicate the presence of PCR biases. As previously shown with PCR amplicon sequencing through pyrotags, chimeras can form and falsely increase microbial diversity (Haas, Gevers et al. 2011). No chimeras were detected in the SEC-TG dataset when using USEARCH version 6.1 (Edgar 2010) indicating SEC does not produce false products that could affect the bacterial diversity.

Utilizing high-throughput sequencing on low biomass environments can unlock the bacterial diversity of these otherwise untapped samples. While SEC avoids PCR amplification and in combination with EMIRGE, recreates full-length 16S rRNA gene sequences, differences were observed between the bacterial families classified with SEC and V4 iTags. While possible chimeras found in V4 iTags could affect the bacterial diversity, this was not further investigated as QPCR on particular families classified was not performed nor were the V4 iTags sequenced in triplicate to further validate the results. SEC did detect *Paenisporosarcina*, a genus found through a clone library and cultivation of the basal ice not detected in V4 iTags. This study demonstrates the successful implementation of SEC in retrieving 16S rRNA genes from a low biomass environment. While large differences were observed in the overall number of bacterial families classified between SEC and V4 iTags, this difference could be attributed to the coverage of bacteria observed with the biotinylated U1406R primer (Chapter 3). Perhaps the use of a primer more selective for bacteria or a combination of the two methods would be the most effective approach when analyzing the bacterial community of low biomass environments.
CHAPTER 6.
SEC AS A TOOL FOR DETECTING INSERTIONS IN AN INSERTIONAL MUTANT LIBRARY OF *CHLAMYDOMAS REINHARDTII*

Introduction

*Chlamydomonas reinhardtii* is a unicellular green alga that serves as a model organism for studies on photosynthesis because of its ability to grow both heterotrophically and autotrophically (Harris 2001). In addition, heterologous DNA can be incorporated into *C. reinhardtii* nuclear, chloroplast, and mitochondrial genomes via transformation (Harris 2001). The frequency of transformation into the nuclear genome of *C. reinhardtii* is approximately $10^3$ events per μg of plasmid DNA when using glass beads (Kindle 1990); this frequency increases to $10^5$ per μg following electroporation (Shimogawara, Fujiwara et al. 1998). The capacity to transform *C. reinhardtii* has been exploited to create insertion mutants (Dent, Haglund et al. 2005, Gonzalez-Ballester, de Montaigu et al. 2005, Zhao, Wang et al. 2009, Barbieri, Larosa et al. 2011), and is the most common means to genetically manipulating this species. Insertion mutants are created by electroporating linear DNA capable of expressing paromomycin resistance into *C. reinhardtii* (Gonzalez-Ballester, de Montaigu et al. 2005), heterologous DNA inserting by an uncharacterized mechanism. Successful transformants are selected with paromomycin. Detection of individual mutants (insertions into a specific locus) requires screening the insertion library using a PCR-based approach (Gonzalez-Ballester, Pootakham et al. 2011). Inserts into genes increase the size of PCR fragments formed from amplicons that flank that gene. A method described by Gonzalez-Ballester et al. (2011) allows investigators to screens thousands of mutants to determine if a gene of interest contains an insertion. Briefly, PCR primers designed to divide the gene of interest into 1 kilobase sections are combined with a primer within the paromomycin resistant gene. Multiple PCR reactions are performed for each
primer combination using genomic DNA isolated from collections of several hundred mutants as a template. A larger fragment band signals that the desired insertion is present in the collection of mutants. The collection is separated into smaller numbers of strains, which are screened in an identical manner until a single isolate containing the insertion is identified. While this method will eventually allow for the isolation of desired mutants, it is a highly laborious process that begins with the assumption that the mutant of interest is present within the insertion library being screened. If it is not present, a great deal of effort and resources can be expended pursuing unproductive screens.

In this study, an attempt was made to use SEC to facilitate the characterization of a *C. reinhardtii* insertion library. The plan was to use SEC to capture insertions by targeting a conserved sequence within the insert expressing paromomycin resistance, capture the adjacent DNA, and use PCR with primers specific for a gene of interest to determine whether an insertion was retrieved. This method could eliminate the need to generate many different primers and perform several PCR reactions per gene of interest, decreasing time and costs.

**Results**

**Locations of Three Insertion Mutants Determined Through SEC**

Three mutants that contained single insertions of the paromomycin resistance cassette were used to establish that SEC could be used to retrieve a portion of the insert and sufficient adjacent DNA to identify the site of the insertion in the *C. reinhardtii* genome. These mutants, designated CAH8-A::AphVIII, CAH8-B::AphVIII, and ATPase::AphVIII, had been previously isolated from an *AphVIII* insertion library created in *C. reinhardtii* (J. V. Moroney, unpublished) and whose locations were determined by the PCR-based method described by Gonzalez-Ballester et al. 2011. A biotinylated capture primer that anneals near the 3´ end of the *AphVIII*
insertion was used to linearly amplify adjacent DNA. SEC was performed in triplicate on each mutant. To determine if there was any carryover of non-targeted genomic DNA, SEC reactions were performed without linear amplification; these controls were termed ‘nopol’ because the DNA polymerase ordinarily added to extend during linear amplification was left out of the reaction. The biotinylated DNA from each SEC reaction and DNA eluted from nopol controls were amplified with primers specific for each insertion, CAH8-A::AphVIII, CAH8-B::AphVIII, and ATPase::AphVIII (Figure 2.2, Figure 2.3). Figure 6.1 demonstrates that SEC successfully captured PCR amplifiable adjacent DNA adjacent to each insertion (see lanes 2, 4, and 6). There was no evidence of amplification in the nopol controls indicating genomic carryover did not contribute to PCR signal obtained following the SEC reactions (Figure 6.1, lanes 3, 5, and 7).

To confirm the PCR products from the SEC reactions belonged to their corresponding mutants, CAH8-A::AphVIII, CAH8-B::AphVIII, and ATPase::AphVIII products were subjected to Sanger sequencing. Sequencing results confirmed the corresponding PCR product was that of each mutant.

Figure 6.1 SEC retrieve DNA adjacent to three loci carrying a paromomycin insert found in three individual insertion mutants of *C. reinhardtii*. Positive PCR amplified SEC reactions are shown in lanes 2, 4, and 6 while corresponding PCR amplified nopol controls remained negative, lanes 3, 5, and 7. Genomic DNA from each individual mutant was PCR amplified and served as positive controls lanes 9, 10, and 11.
Capture of a Single Insertion From 180 Mutants Through SEC

Since SEC was effective in retrieving adjacent DNA from individual insertions (Figure 6.1), we attempted to retrieve a single insertion mutant within a pool of 180 insertion mutants. A 180 pool was created that combined the ATPase::AphVIII mutant with 179 paromomycin strains randomly chosen from the C. reinhardtii insertion library. Two SEC reactions (SEC-180 pool 1 and SEC-180 pool 2) were performed on genomic DNA extracted from this 180 pool, and QPCR was performed in triplicate on each SEC reaction to quantify the number of ATPase::AphVIII inserts present. To determine how much non-targeted genomic carryover was present, quantitative PCR was performed using the single copied gene cblp. SEC successfully captured DNA adjacent to the ATPase::AphVIII insert above background (Table 6.1). In addition, nopol controls were performed in triplicate to further assess any genomic DNA carryover. In all nopol control reactions, genomic DNA carryover was below SEC recoveries, providing evidence to support that we could retrieve and identify a single insertion from within a mix of 180 insertion mutants. Based on this result, we assumed that SEC could retrieve the remaining 179 mutants with equal efficiency.

Table 6.1 SEC capture of a single insertion mutant, ATPase::AphVIII from 180 pool.

<table>
<thead>
<tr>
<th>SEC reactions</th>
<th>Molecules of ATPase::AphVIII (x10^3)</th>
<th>Molecules of ATPase::AphVIII (x10^3) nopol control</th>
<th>Molecules of cblp gene (x10^3)</th>
<th>Molecules of cblp gene (x10^3) nopol control</th>
</tr>
</thead>
<tbody>
<tr>
<td>SEC-180 pool 1</td>
<td>390 ± 32</td>
<td>1.0 ± 0.8</td>
<td>0.8 ± 0.02</td>
<td>4.6 ± 1.4</td>
</tr>
<tr>
<td>SEC-180 pool 2</td>
<td>600 ± 27</td>
<td>3.5 ± 1.4</td>
<td>0.4 ± 0.05</td>
<td>3.9 ± 2.4</td>
</tr>
<tr>
<td>Mean</td>
<td>500 ± 120</td>
<td>2.2 ± 1.7</td>
<td>12 ± 4.0</td>
<td>4.2 ± 1.8</td>
</tr>
</tbody>
</table>

Note: Numbers are means ± standard deviation.
Capture of A Single Insertion from 1,440 Mutants (1,440 Pool) Through SEC

To explore the limits of SEC in this system, an attempt was made to retrieve a single insertion mutant within a pool created by 1,440 insertion mutants. Equal amounts of genomic DNA from eight different 180 pools were combined together to generate the 1,440 pool. A single ATPase::AphVIII insertion was present in one of the 180 pools. SEC was performed and QPCR was used to determine the number of ATPase::AphVIII insertion retrieved. SEC was effective in capturing the ATPase mutant in the two SEC reactions attempted and the number of molecules retrieved was found above background in those reactions (Table 6.2). The ATPase::AphVIII insertion could not be detected in the nopol reactions, values being below the Ct threshold necessary for estimating quantity by QPCR. The cblp gene, which estimates non-targeted DNA carryover, was present at levels three- to ten-fold lower than the ATPase::AphVIII insertion.

Table 6.2 SEC captures insertion mutants from the 1,440 mix pool sample that contained over 1,440 mutants from an insertion library.

<table>
<thead>
<tr>
<th>SEC reactions</th>
<th>Molecules of ATPase::AphVIII (x10³)</th>
<th>Molecules of ATPase::AphVIII (x10³) nopol control</th>
<th>Molecules of cblp gene (x10³)</th>
<th>Molecules of cblp gene (x10³) nopol control</th>
</tr>
</thead>
<tbody>
<tr>
<td>1,440 pool 1</td>
<td>2.2 ± 3.3</td>
<td>BT</td>
<td>0.93 ± 0.50</td>
<td>1.4 ± 2.4</td>
</tr>
<tr>
<td>1,440 pool 2</td>
<td>1.4 ± 0.46</td>
<td>BT</td>
<td>BT</td>
<td>11 ± 10</td>
</tr>
<tr>
<td>1,440 pool 3</td>
<td>4.6 ± 2.1</td>
<td>BT</td>
<td>0.43 ± 0.25</td>
<td>BT</td>
</tr>
<tr>
<td>Mean</td>
<td>2.8 ± 2.5</td>
<td>BT</td>
<td>0.7 ± 0.4</td>
<td>5.5 ± 8.0</td>
</tr>
</tbody>
</table>

Note: Numbers are shown as means ± the standard deviation. BT = below Ct threshold.
Discussion

*C. reinhardtii* is a model organism for studies of photosynthesis and respiration because of the organism’s ability to grow both autotrophically and heterotrophically (Harris 2001) allowing for the generation of mutants required for both autotrophy and heterotrophy. While mutants are often generated in large numbers using a reverse genetics approach through electroporation and the integration of linearized DNA (Gonzalez-Ballester, Pootakham et al. 2011), determining the location of these insertions is a labor-intensive process involving multiple rounds of PCR amplification. In this study, SEC was used to capture the end of a single insertion and adjacent DNA in an effort to quickly and accurately determine whether individual insertions can be found within an insertion library.

The biotinylated capture primer used for SEC was located 180 bases upstream of the 3´ end of the insertion sequence. SEC was tested on three individual insertions and the protocol was effective in capturing these sequences. The identity of individual insertions was confirmed by PCR amplification and Sanger sequencing of DNA adjacent to the insertion site of three mutants (Figure 6.1). SEC’s capacity to retrieve an individual insertion was tested further by attempting to retrieve a single insertion from a pool of 180 and 1,440 insertion mutants. In both circumstances, SEC recovered the targeted insertion and did so without evidence of non-specific carryover of genomic DNA. For both pools of insertion mutant DNA, the presence of the non-targeted *cblp* gene in the reaction was lower than that of the targeted sequence.

The next logical step would be using SEC to assess the locations of a large number of insertions in an insertion library pool at one time. To do this, one would perform SEC on a pool of insertion mutants and combine with high-throughput sequencing thereby sequencing the adjacent DNA to determine the locations of a large number of the insertions retrieved. This
would allow for the detection of multiple insertions without the highly laborious rounds of PCR amplification, a method currently used to determine if an insertion is present in a single locus. However, a number of issues need to be addressed before implementing SEC in this manner. Insertion libraries generated through electroporation of *C. reinhardtii* have given rise to many challenges making it difficult to determine where an insertion is located using the PCR-based method described by Gonzalez-Ballester et al. (2011). It was found that upon transformation to generate insertion mutants, multiple insertions were found in the same gene with a duplication of genomic DNA (Aksoy, Pootakham et al. 2013). Insertion mutants may also contain deletions. *C. reinhardtii* has been shown to delete portions of genomic DNA surrounding insertion sites by as much as 35,000 base pairs (Dent, Haglund et al. 2005, Gonzalez-Ballester, Pootakham et al. 2011), which makes determining which gene responsible for a particular phenotype extremely difficult. SEC allows for capture of adjacent DNA from only a single end, deletions at the other end of the insertion will go unnoticed; the capture of adjacent DNA in both directions would be beneficial by capturing both ends of the insertion. This problem is further complicated by the ability of *C. reinhardtii* to truncate insertion sequences at either end of the insertion (Gonzalez-Ballester, Pootakham et al. 2011); truncation where the biotinylated primer would be located would result in failure to capture. In addition, the sheer size of the *C. reinhardtii* genome, over 113 Mb, could allow for potential non-targeted amplification using the current biotinylated primer. In example, the last ten bases of the biotinylated capture primer sequence appears 93 times in the genome. Perhaps a longer biotinylated capture primer than the one used in this study could alleviate this issue. Alternatively, the addition of a barcode sequence in the insertion sequence that is not found in the *C. reinhardtii* genome could serve as an annealing site for the biotinylated capture primer to avoid non-targeted amplification.
Overall, SEC can be used to determine the whether a gene of interest contains an insertion in the presence of 1,440 insertion mutants. SEC would be performed and the biotinylated product could be PCR amplified with primers specific for that gene. This method decreases the amount of reagents and primers used when compared to method outlined by Gonzalez-Ballester et al. 2011. The development of a more specific biotinylated primer and the development of a bioinformatic pipeline to locate insertion site are necessary for application of SEC and high-throughput sequencing for retrieving all insertion mutants from a larger pool of insertions.
CHAPTER 7.
CONCLUDING REMARKS AND FUTURE DIRECTIONS

The ability to enrich and capture regions of a genome with little prior knowledge of the composition of those regions offer opportunity to answer specific questions without bias. The main goal of this research was to develop a method that would decrease biases associated with amplicon sequencing by capturing and retrieving specific DNA sequences from complex mixtures, allowing for enrichment without PCR and provide an alternative method to characterizing bacteria from environmental communities.

SEC was shown capable of capturing the gltS gene, along with 5,000 bases of adjacent DNA (Chapter 3). A previously developed method that used biotinylated oligonucleotides obtained lengths of only 1,000 base pairs (Sterky, Holmberg et al. 1998), and to date no capture method have obtained biotinylated product of the lengths reported here. Given the length of captured biotinylated DNA achievable through SEC, one could use this method as a way to close genomes being sequenced that contain gaps. This could be achieved by designing biotinylated capture primers using known sequences on either end of the gap, performing SEC, and sequencing the biotinylated DNA product. Potentially larger capture fragments could be achieved through potential modification of SEC protocol. While steps were taken to avoid shearing – use of wide-born pipet tips – adjustments of the concentrations of DNA polymerase, dNTPs, and linear amplification cycle times could increase capture fragment lengths.

The utility of SEC is not limited to a single gene in one organism. Loci that are conserved in all or a group of members in a community could be captured and the diversity of that gene could be obtained. The 16S rRNA gene, present in all bacteria, was targeted and captured from an artificial bacterial mock community with twenty individuals of varying abundances (Chapter 3), allowing an accurate assessment of bacterial diversity within that
In addition, SEC was compatible with high-throughput sequencing using the Illumina HiSeq 2000 platform and outperformed PCR amplicon sequencing in retrieving sequence from species of lower abundance within the mock community (Chapter 3). SEC was also successful in the capture of 16S rRNA genes from three marine samples, seawater, sediment, and within the Eastern oyster, *C. virginica* (Chapter 4). SEC offers a method to determine the bacterial diversity within a eukaryotic host in the overabundance of eukaryotic DNA. The ability to determine bacterial diversity within eukaryotic hosts can help evaluate these communities containing an excess of eukaryotic DNA and provide an alternative method to conventional metagenomic and PCR amplicon approaches to sequencing these environments.

While PCR amplification prior to library preparations was avoided, PCR amplification was still a necessary requirement of Illumina sequencing platforms for the addition of Illumina adapters and as a result potential biases could have been introduced that were not investigated. Further work is necessary to determine whether the PCR amplification during library preparation could have introduced biases. PCR-free library preparations for Illumina sequencing that were not available at the time these studies were performed are now available and performed simultaneously with PCR amplified SEC capture fragments could provide evidence of potential biases being introduced.

Chapter 5 displayed the ability of SEC to enrich 16S rRNA genes from a low biomass environment, the basal ice from Taylor Glacier, Antarctica with a DNA concentration of 21 pg/μL (Doyle, Montross et al. 2013). However, bacterial family classification between SEC and V4 iTags differed. A limitation in the use of SEC is the selection of the biotynilated capture primer. As shown in Chapter 3, the capture primer used for retrieval of the 16S rRNA gene matched to 69.2% of bacteria (Klindworth, Pruesse et al. 2013) potentially missing bacteria.
present in those environments. Alternative primers targeting the 16S rRNA gene are available with more specificity than the biotinylated Uni1406R used in these studies.

Another caveat to using SEC is the limitations with EMIRGE. The program EMIRGE is necessary to reconstruct full-length 16S rRNA genes from the SEC-derived sequencing data. Limitations with EMIRGE prevent reconstruction of 16S rRNA gene sequences if those sequences do not contain sufficient coverage and depth of the gene. Access to using more the SEC-derived dataset for EMIRGE analysis could overcome this problem and would require special computational considerations.

SEC is also not limited solely to bacterial 16S rRNA gene. Klinderworth et al. 2013 provides primers specific to Archaea and Eukarya. Additionally, other loci of interest that are conserved in all members of the community could be captured and the diversity of those genes could be obtained. A specific gene found in groups of organisms can be targeted, for example, the addition of a biotin molecule to a methanogen-specific primer can target and retrieve that sequence from an environmental sample.

SEC is not limited for implementation in just bacterial genomes. The SEC protocol was demonstrated to be successful in a eukaryotic system, the unicellular alga *Chlamydomonas reinhardtii* (Chapter 6) by capturing DNA adjacent to individual insertion mutants in an insertion library. While more work in necessary to increase the utility of SEC with a larger number of insertion mutants, the work provided in this thesis provides the groundwork for SEC’s use in targeting a gene in a eukaryote. Other applications for SEC could be to capture conserved motifs in protein structures using the genomic DNA. A biotinylated primer could be designed using nucleotide sequence information of the conserved motif. DNA isolated from an environment would then undergo SEC, capturing all the sequences that matched that motif. One could then
sequence the biotinylated DNA via high-throughput sequencing, to observe single nucleotide polymorphisms of the motif. Either a single or multiple DNA regions could be targeted and concentrated at one time within a given sample. SEC can also be easily multiplexed with the capture of multiple loci by providing a barcode to the 5´ end of the biotinylated capture primer. Post SEC, the biotinylated product from each locus captured could be sequenced to obtain high coverage of all regions. Finally, this method could be used to observe microbial community changes over time in response to an environmental change by retrieving phylogenetic marker genes, such as the 16S rRNA gene. Samples could be multiplexed together to provide several snap-shots of the community in response to a particular environmental change. The ability of SEC to capture specific loci from artificial metagenomes, complex metagenomes in the overabundance of eukaryotic DNA, and low biomass environments showcase the utility of SEC to work in a variety of environments.
LITERATURE CITED


APPENDIX

TABLE 1

Table 1. Alignment of S-⁎-Univ-1390-a-A-18 (Uni1390R) targeting position 1407 → 1390 of the 16S rRNA gene of *E. coli* with all 16S rRNA genes found in the HM-277D metagenome. Complementary bases are underlined.

Acinetobacter baumannii ATCC 17978
GGCC\textcolor{green}{TTGTACACACCCGCCCCTCAACCC}  
GGCC\textcolor{green}{TTGTACACACCGCCCGTCACACC}  
GGCC\textcolor{green}{TTGTACACACCGCCCGTCACACC}  
GGCC\textcolor{green}{TTGTACACACCGCCCGTCACACC}  
GGCC\textcolor{green}{TTGTACACACCGCCCGTCACACC}

Actinomyces odontolyticus, strain 1A.21
GGCC\textcolor{green}{TTGTACACACCCGCCCCTCAACCGTC}  
GGCC\textcolor{green}{TTGTACACACCGCCCGTCACACC}  

Bacillus cereus, strain NRS 248 ATCC 10987
GGCC\textcolor{green}{TTGTACACACCGCCCGTCACACC}  
GGCC\textcolor{green}{TTGTACACACCGCCCGTCACACC}  
GGCC\textcolor{green}{TTGTACACACCGCCCGTCACACC}  
GGCC\textcolor{green}{TTGTACACACCGCCCGTCACACC}  
GGCC\textcolor{green}{TTGTACACACCGCCCGTCACACC}

Bacteroides vulgatus, strain NCTC 11154 ATCC 8482
GGCC\textcolor{green}{TTGTACACACCGCCCGTCACACC}  
GGCC\textcolor{green}{TTGTACACACCGCCCGTCACACC}  
GGCC\textcolor{green}{TTGTACACACCGCCCGTCACACC}  
GGCC\textcolor{green}{TTGTACACACCGCCCGTCACACC}  
GGCC\textcolor{green}{TTGTACACACCGCCCGTCACACC}

Clostridium beijerinckii, strain NCIMB 8052
GGCC\textcolor{green}{TTGTACACACCGCCCGTCACACC}  
GGCC\textcolor{green}{TTGTACACACCGCCCGTCACACC}  
GGCC\textcolor{green}{TTGTACACACCGCCCGTCACACC}  
GGCC\textcolor{green}{TTGTACACACCGCCCGTCACACC}  
GGCC\textcolor{green}{TTGTACACACCGCCCGTCACACC}
Deinococcus radiodurans, strain R1 ATCC13939
GGCCTTGTAACACCGCCCGTCACACC
GGCCTTGTAACACCGCCCGTCACACC
GGCCTTGTAACACCGCCCGTCACACC
GGCCTTGTAACACCGCCCGTCACACC
GGCCTTGTAACACCGCCCGTCACACC
GGCCTTGTAACACCGCCCGTCACACC

Enterococcus faecalis OG1RF
GGCCTTGTAACACCGCCCGTCACACC
GGCCTTGTAACACCGCCCGTCACACC
GGCCTTGTAACACCGCCCGTCACACC
GGCCTTGTAACACCGCCCGTCACACC
GGCCTTGTAACACCGCCCGTCACACC

Escherichia coli, strain K12, MG1655
GGCCTTGTAACACCGCCCGTCACACC
GGCCTTGTAACACCGCCCGTCACACC
GGCCTTGTAACACCGCCCGTCACACC
GGCCTTGTAACACCGCCCGTCACACC
GGCCTTGTAACACCGCCCGTCACACC

Helicobacter pylori, strain 26695
GGTCCTGTACTCACGCCCCGTCACACC
GGTCCTGTACTCACGCCCCGTCACACC

Lactobacillus gasseri ATCC 33323
GGCCTTGTAACACCGCCCGTCACACC
GGCCTTGTAACACCGCCCGTCACACC
GGCCTTGTAACACCGCCCGTCACACC
GGCCTTGTAACACCGCCCGTCACACC
GGCCTTGTAACACCGCCCGTCACACC

Listeria monocytogenes EGD-e
GGCCTTGTAACACCGCCCGTCACACC
GGCCTTGTAACACCGCCCGTCACACC
GGCCTTGTAACACCGCCCGTCACACC
GGCCTTGTAACACCCGCCCTTAACCC
GGCCTTGTAACACCCGCCCTTAACCC
GGCCTTGTAACACCCGCCCTTAACCC

Neisseria meningitidis MC58
GGTCCTTGTAACACCCGCCCTTAACCC
GGTCCTTGTAACACCCGCCCTTAACCC
GGTCCTTGTAACACCCGCCCTTAACCC
GGTCCTTGTAACACCCGCCCTTAACCC
VITA

Kelley G. Núñez, a native of Kenner, Louisiana, was born to Ralph and Kathleen in 1986. Kelley graduated from Alfred Bonnabel High School in Kenner, Louisiana in 2005. In May 2009, Kelley earned a Bachelor of Science degree in Microbiology with a minor in Chemistry from Louisiana State University in Baton Rouge, Louisiana. She remained at Louisiana State University to pursue a Doctorate of Philosophy degree in Biological Sciences. She is married to Kevin Núñez and they have a son named Keegan Michael Núñez.