POSSIBLE DETECTION OF PATHOGENIC BACTERIAL SPECIES INHABITING STREAMS IN GREAT SMOKY MOUNTAINS NATIONAL PARK

A thesis presented to the faculty of the Graduate School of Western Carolina University in partial fulfillment of the requirement for the degree of Master of Science in Biology.

By

Kwame Agyapong Brown

Director: Dr. Sean O’Connell
Associate Professor of Biology and
Biology Department Head

Committee members: Dr. Sabine Rundle and Dr. Timothy Driscoll

November 2016
ACKNOWLEDGEMENTS

I would like to express my deepest gratitude to my adviser Dr. Sean O’Connell for his insightful mentoring and thoughtful contribution throughout this research. I would also like to acknowledge my laboratory teammates: Lisa Dye, Rob McKinnon, Kacie Fraser, and Tori Carlson for their diverse contributions to this project. This project would not have been possible without the generous support of the Western Carolina University Biology Department and stockroom; so I would like to thank the entire WCU biology faculty and Wesley W. Bintz for supporting me throughout my Masters program. Finally, I would like to thank my thesis committee members and reader: Dr. Sabine Rundle, Dr. Timothy Driscoll and Dr. Anjana Sharma for their contributions to this project.
# TABLE OF CONTENTS

List of Tables ................................................................................................................................. iv  
List of Figures ..................................................................................................................................v  
Abstract ........................................................................................................................................ vii  
Introduction ......................................................................................................................................1  
 Specific Aims ..................................................................................................................................3  
 Significance ...................................................................................................................................3  
Methods ...........................................................................................................................................5  
 Site Description .............................................................................................................................5  
 Sample Collection and Storage ....................................................................................................5  
 Recovery of Cells from Capsules .................................................................................................5  
 Enrichment Culturing and Screening of Bacteria *(Gram Negative vs. Gram Positive)* ............7  
 Pathogen Selection .........................................................................................................................8  
 Selection for Pseudomonas Species ...............................................................................................9  
 Selection for Salmonella typhimurium .........................................................................................9  
 Selection for Gram negative Enteric Bacilli ..............................................................................9  
 Selection for Staphylococcal Species..........................................................................................10  
 Isolate Designations .....................................................................................................................12  
 DNA Sequencing and Culture Matching .....................................................................................12  
 Phylogenetic Analysis ...................................................................................................................13  
Results ...........................................................................................................................................15  
 Classification of Isolates .............................................................................................................15  
 SeqMatch Results & Phylogenetic Trees .....................................................................................15  
Discussion .......................................................................................................................................29  
 Isolate Selection and Classification ...........................................................................................29  
Conclusion ......................................................................................................................................35  
References ......................................................................................................................................38  
Appendix .........................................................................................................................................42  
 Water Pump Description ..............................................................................................................42  
 16S rDNA Sequences.....................................................................................................................43
LIST OF TABLES

Table 1. List of four sites sampled for bacteria in Great Smoky Mountains National Park including GPS coordinates and the volume of water filtered (L) at each site .................................6
Table 2. List of media used for enrichment and pathogen selection from four streams in Great Smoky Mountains National Park. ........................................................................................................8
Table 3. RDP Classifier results showing the genus each isolate obtained from GSMNP belonged to, and the accuracy (confidence) of the result in percentage.............................................................15
Table 4. List of type strain bacterial species with ≥ 90% 16S rDNA sequence match to isolate NCKP1 and their percent identity to NCKP1 ........................................................................................................16
Table 5. List of type strain bacterial species with ≥ 90% 16S rDNA sequence match to isolates TLPR1 and NCOR2 and their percent identity to TLPR1 and NCOR2 ........................................18
Table 6. List of type strain bacterial species with ≥ 90% 16S rDNA sequence match to isolate NCOR1 and their percent identity to NCOR1 ..................................................................................21
Table 7. List of type strain bacterial species with ≥ 90% 16S rDNA sequence match to isolate NCOR3 and their percent identity to NCOR3 ..................................................................................21
Table 8. List of type strain bacterial species with ≥ 90% 16S rDNA sequence match to isolates NCKP2 and THKB2 and their percent identity to NCKP2 and THKB2 .................................23
Table 9. List of type strain bacterial species with ≥ 90% 16S rDNA sequence match to isolate THKB1 and their percent identity to THKB1 ...................................................................................25
Table 10. List of type strain bacterial species with ≥ 90% 16S rDNA sequence match to isolates TLPR1 and TLPR2 and their percent identity to TLPR1 and TLPR2 .................................................27
LIST OF FIGURES

Figure 1: Schematic showing selective protocol sequence used to isolate potentially pathogenic bacterial species from water samples collected from four streams in Great Smoky Mountains National Park and the media type on which each isolate was isolated (key for isolates: NC = North Carolina, T = Tennessee; KP = Kephart Prong, OR = Oconaluftee River, HKB = Hickory King Branch, LPR = Little Pigeon River; 1, 2, 3 = sequential isolates obtained from each site...11

Figure 2a: Phylogenetic tree constructed using the top 12 sequence matches from the Ribosomal Database Project (type & non-type strains) analyzed with MUSCLE multiple sequence alignment tool, Gblocks selective tool, PhyML phylogeny tools and rendered via TreeDyn tool for isolate NCKP1 from Great Smoky Mountains National Park.................................................................16

Figure 2b: Phylogenetic tree constructed using the top 12 sequence matches from the Ribosomal Database Project (type & non-type strains), and analyzed with MUSCLE multiple sequence alignment tool, Gblocks selective tool, PhyML phylogeny tool and rendered via TreeDyn tool with branch distance ignored for isolate NCKP1 from Great Smoky Mountains National Park. Branch support values are out of 1.0. .................................................................17

Figure 3a: Phylogenetic tree constructed using the top 9 sequence matches from the Ribosomal Database Project (type & non-type strains), and analyzed with MUSCLE multiple sequence alignment tool, Gblocks selective tool, PhyML phylogeny tool and rendered via TreeDyn tool for isolates NCOR2 and TLPR1 from Great Smoky Mountains National Park..............................18

Figure 3b: Phylogenetic tree constructed using the top 9 sequence matches from the Ribosomal Database Project (type & non-type strains), and analyzed with MUSCLE multiple sequence alignment tool, Gblocks selective tool, PhyML phylogeny tool and rendered via TreeDyn tool with branch distance ignored for isolates NCOR2 & TLPR1 from Great Smoky Mountains National Park. Branch support values are out of 1.0 ..................................................................19

Figure 4a: Phylogenetic tree constructed using the top 9 sequence matches from the Ribosomal Database Project (type & non-type strains), and analyzed with MUSCLE multiple sequence alignment tool, Gblocks selective tool, PhyML phylogeny tool and rendered via TreeDyn tool for isolate NCOR1 from Great Smoky Mountains National Park .......................................................20

Figure 4b: Phylogenetic tree constructed using the top 9 sequence matches from the Ribosomal Database Project (type & non-type strains), and analyzed with MUSCLE multiple sequence alignment tool, Gblocks selective tool, PhyML phylogeny tool and rendered via TreeDyn tool with branch distance ignored for isolate NCOR1 from Great Smoky Mountains National Park. Branch support values are out of 1.0 .................................................................20

Figure 5a: Phylogenetic tree constructed using the top 8 sequence matches from the Ribosomal Database Project (type & non-type strains), and analyzed with MUSCLE multiple sequence alignment tool, blocks selective tool, PhyML phylogeny tool and rendered via TreeDyn tool for isolate NCOR3 from Great Smoky Mountains National Park.................................................................21

Figure 5b: Phylogenetic tree constructed using the top 8 sequence matches from the Ribosomal Database Project (type & non-type strains), and analyzed with MUSCLE multiple sequence alignment tool, Gblocks selective tool, PhyML phylogeny tool and rendered via TreeDyn tool for isolate NCOR3 from Great Smoky Mountains National Park. Branch support values are out of 1.0 ..................................................................22
Figure 6a: Phylogenetic tree constructed using the top 8 sequence matches from the Ribosomal Database Project (type & non-type strains), and analyzed with MUSCLE multiple sequence alignment tool, Gblocks selective tool, PhyML phylogeny tool and rendered via TreeDyn tool for isolates THKB2 & NCKP2 from Great Smoky Mountains National Park..........................................................24
Figure 6b: Phylogenetic tree constructed using the top 8 sequence matches from the Ribosomal Database Project (type & non-type strains), and analyzed with MUSCLE multiple sequence alignment tool, Gblocks selective tool, PhyML phylogeny tool and rendered via TreeDyn tool with branch distance ignored for isolates THKB2 & NCKP2 from Great Smoky Mountains National Park. Branch support values are out of 1.0 ..............................................................................................24
Figure 7a: Phylogenetic tree constructed using the top 8 sequence matches from the Ribosomal Database Project (type & non-type strains), and analyzed with MUSCLE multiple sequence alignment tool, blocks selective tool, PhyML phylogeny tool and rendered via TreeDyn tool for isolate THKB1 from Great Smoky Mountains National Park..........................................................25
Figure 7b: Phylogenetic tree constructed using the top 8 sequence matches from the Ribosomal Database Project (type & non-type strains), and analyzed with MUSCLE multiple sequence alignment tool, Gblocks selective tool, PhyML phylogeny tool and rendered via TreeDyn tool with branch distance ignored for isolate THKB1 from Great Smoky Mountains National Park. Branch support values are out of 1.0 ..............................................................................................26
Figure 8a: Phylogenetic tree constructed using the top 8 sequence matches from the Ribosomal Database Project (type & non-type strains), and analyzed with MUSCLE multiple sequence alignment tool, Gblocks selective tool, PhyML phylogeny tool and rendered via TreeDyn tool for isolates TLPR2 & TLPR3 from Great Smoky Mountains National Park..........................................................27
Figure 8b: Phylogenetic tree constructed using the top 8 sequence matches from the Ribosomal Database Project (type & non-type strains), and analyzed with MUSCLE multiple sequence alignment tool, Gblocks selective tool, PhyML phylogeny tool and rendered via TreeDyn tool with branch distance ignored for isolates TLPR2 & TLPR3 from Great Smoky Mountains National Park. Branch support values are out of 1.0 ..............................................................................................28
POSSIBLE DETECTION OF PATHOGENIC BACTERIAL SPECIES INHABITING STREAMS IN GREAT SMOKY MOUNTAINS NATIONAL PARK

Kwame Agyapong Brown, M.S.
Western Carolina University, November 2016
Director: Dr. Sean O’Connell

Numerous pathogenic bacterial species have been found in many freshwater systems around the world. These pathogens affect the overall water quality of these systems and may cause diseases in both aquatic and terrestrial animals which may lead to loss of species diversity and abundance in their environments. This study sought to identify and document pathogenic bacterial species that may inhabit the streams that flow through Great Smoky Mountains National Park. Bacterial cells were collected by filtering water from four streams (Oconaluftee River, Kephart Prong, Little Pigeon River and Hickory King Branch Stream) through separate capsule filters. The cells were later backflushed from the filters and cultured on various selective and differential media. Ten isolates were selected based on phenotypic characteristics such as colony color and growth on specific media type, and sample origin. The nearly full 16S rDNA was sequenced for all ten isolates and analyzed to determine their identity.

Out of the ten isolates, four isolates were from the phylum Firmicutes while the other six were in the phylum Proteobacteria. Phylogenetic analysis of these isolates showed eight out of the ten isolates were related to known opportunistic pathogens. The other two were related to a
ubiquitous *Bacillus* species that is considered to be a probiotic. Although none of the isolates had a 100% match to a known obligate or opportunistic pathogen, many isolates matched > 97% to opportunistically pathogenic species. Follow up molecular and metabolic tests need to be employed to determine the pathogenicity of each isolate.
INTRODUCTION

Pristine water resources, both surface and groundwater, are becoming more scarce because of global increases in population and the action of humans in the environment (Geldreich 1996). Water-borne pathogen contamination in water resources and related diseases are a major water quality concern throughout the world (Pandey et al. 2014).

Bacteria and other single-celled microbes play a vital role in freshwater ecosystems (Findlay 2010). They perform numerous processes such as organic matter breakdown and nitrogen fixation (Findlay 2010). Bacteria are essential to the health of all water systems and also serve as food for other organisms present in the water bodies (Cairns 1971). Conversely, some bacterial species (classified as pathogens) are harmful to aquatic and other organisms that share the same freshwater body, including terrestrial organisms that drink the water.

Numerous pathogenic bacterial species have been found in many freshwater systems around the world (Geldreich 1996). These pathogens affect the overall water quality of these systems (Pandey et al. 2014) and cause diseases in both aquatic and terrestrial animals which may lead to loss of species diversity or species abundances in their environments. For example *Clostridium botulinum* and *Vibrio cholerae* are common pathogenic bacteria that cause fatal illnesses in terrestrial animals that drink water that is contaminated with fecal matter (Cabral 2010). Other pathogenic bacteria such as *Legionella pneumophila* are opportunistic and cause diseases in both aquatic and terrestrial animals that are immunocompromised or have a break in their physical protective barrier such as a cut on their skin (Dey 2009).

Water quality is routinely evaluated by determining the concentration of coliforms such as *Escherichia coli* (American Public Health Association 1965). Although this method is an
efficient way of determining water quality based on a particular subset of potential pathogenic bacteria, it does not identify the different species of bacterial pathogens that are present in the water and misses many other types.

Great Smoky Mountains National Park (GSMNP) contains one of the most diverse assemblages of plants, animals, microbes and invertebrates species in North America (Jenkins 2007). According to the National Park Service (NPS) only 18,545 species have been documented in the park: Scientists believe an additional 30,000-80,000 species of plants and animals (excluding bacteria and other microorganisms) may live there. Also there are about 19 major streams that exist in the park and these streams serve as a drinking water source for many of the diverse animals found in the park.

In 1998, GSMNP embarked on a project to determine all life forms that resided or spent time in the Park (Nichols & Langdon 2007). By 2006 over 300 species of bacteria and archaea were identified using DNA sequencing technology; many of them are novel species that have never before been described (Nichols & Langdon 2007). There is still ongoing research by systematists to identify the estimated 100,000 species of organisms that are found in the GSMNP, both microbial and larger.

Previous research performed by O’Connell et al. (2007) to inventory bacteria found within the soils and streams of the GSMNP revealed there was a total of eleven phyla present including six which were found only via culture-independent techniques. Overall they found about 69 genera of bacteria with Bacteroidetes being the dominant phylum in water (O’Connell et al. 2007). Although this research focused on general bacteria diversity and did not select for known pathogenic bacterial species, results from the study alluded to the presence of pathogenic species in the streams sampled. Bacteria from the genera Bacillus, Pseudomonas, Enterobacter
and Staphylococcus were cultured from water sampled from GSMNP (O’Connell et al. 2007). Certain species from these genera such as Bacillus anthracis and Pseudomonas aeruginosa are clinically relevant to humans, livestock and wildlife.

There is little known about the diversity of water-borne pathogenic bacteria that exist in the streams in the park. Since there are hundreds of water-borne pathogenic bacteria that can exist in a stream (Geldreich 1996), this study focused on those that could cause disease in terrestrial mammalian species that inhabit the Great Smoky Mountains including but not limited to Staphylococcus, Pseudomonas and Salmonella species. With the aid of 16S ribosomal DNA gene sequencing techniques and phylogenetic analysis, pathogenic bacterial species were targeted to be identified (Emerson et al. 2008), and added to the All Taxa Biodiversity Inventory (ATBI; Nichols and Langdon 2007). For this study, a 100% DNA sequence match to a known type strain pathogenic bacteria that shared an immediate common ancestor with the same type strain species would be counted as a positive finding of a pathogen from GSMNP. This work will lay the groundwork for further studies that will confirm each isolate’s pathogenicity via immunoassays and antibiotic resistance testing, for example.

Specific Aims

The purpose of this research was to sample, identify and document potential pathogenic bacterial species that may inhabit the Oconaluftee River, Kephart Prong, Little Pigeon River and Hickory King Branch.

Significance

This study will help assess the water quality in GSMNP streams. Poor water quality could lead to various diseases that could end up sickening or killing many of the animals in the park that rely on these streams. In addition, knowing the diversity of pathogenic bacteria in these
streams could help the NPS to develop better management systems to control proliferation of these species and maintain pristine water quality.
METHODS

Site Description

Four streams were selected from GSMNP; two of the streams were located in North Carolina while the other two were located in Tennessee (Table 1). In North Carolina the streams sampled were the Oconaluftee River and Kephart Prong; and in Tennessee the streams sampled were the Little Pigeon River and Hickory King Branch. The Oconaluftee and Little Pigeon rivers are large streams (>18 ft wide) with little plant cover over the middle, but the banks and edges are surrounded by trees and other vegetation. The Little Pigeon River is rockier and faster flowing than the Oconaluftee River which is slow moving. Kephart Prong and Hickory King Branch are at a higher elevation and smaller than the other two streams (<18 ft wide). Kephart Prong is rocky and fast flowing while Hickory King Branch flowed slower than the other three. Both Hickory King Branch and Kephart Prong had a greater degree of shade compared to the Little Pigeon River and the Oconaluftee River, with Hickory King Branch being the most shaded among the sites sampled.

Sample Collection and Storage

Water samples were collected from four different sites using aseptic techniques. A custom-made water pump filtration system (description in Appendix) was used to draw water from the streams through a 0.22 μm sterile capsule filter (Pall Gelman Sciences). A 6-gallon Jerry can was used to estimate the amount of water filtered from each sample site. Each capsule was clamped shut using a small length of tubing and sealed after the designated amount of water had been filtered and they were then immediately stored on ice in a cooler at approximately 4°C. The capsules were later refrigerated at 4°C in the laboratory for 12 days.
Table 1: List of four sites sampled for bacteria in Great Smoky Mountains National Park including GPS coordinates and the volume of water filtered (L) at each site.

<table>
<thead>
<tr>
<th>State</th>
<th>Sample Site</th>
<th>GPS Coordinates</th>
<th>Volume of Water Filtered (L)</th>
</tr>
</thead>
<tbody>
<tr>
<td>North Carolina</td>
<td>Kephart Prong (NCKP)</td>
<td>35.35.222 N 83.21.431 W</td>
<td>22.7</td>
</tr>
<tr>
<td></td>
<td>Oconaluftee (NCOR)</td>
<td>35.30.710 N 83.81.150 W</td>
<td>34.1</td>
</tr>
<tr>
<td>Tennessee</td>
<td>Little Pigeon River (TLPR)</td>
<td>35.63.336 N 83.51.526 W</td>
<td>34.1</td>
</tr>
<tr>
<td></td>
<td>Hickory King Branch (THKB)</td>
<td>35.68.700 N 85.53.484 W</td>
<td>22.7</td>
</tr>
</tbody>
</table>

Recovery of Cells from Capsules

Bacterial cells were backflushed from the filter capsules. The backflushing protocol used to recover cells from the 0.22 μm filter involved warming the filters gradually to room temperature. Once each filter was brought to room temperature, ~150 mL of 0.1% sodium pyrophosphate (NaPP) was added (~80 mL into the outlet end and ~70 mL to the inlet end of the filter capsule). The filter ends were sealed using sterile tubing attached tightly to the filter and clamped shut. The filters were shaken on a culture shaker for 5 minutes at 400 rpm and then the clamp was removed on one end and sterile plastic tubing (tubing was sterilized in 10% bleach solution for five minutes and rinsed with sterile deionized water and 0.1% NaPP) was attached. This was then repeated for the other end of the capsule. The free end of tubing attached to the outlet end of the filter housing was tightly connected to a pressurized air nozzle in the laboratory. Air was forced into the filter housing at low to medium pressure and discharge collected from the inlet end in order to remove cells from the filter material. A total of approximately 150mL of
discharge was collected from the capsule and placed in a 1000-ml Erlenmeyer flask. This protocol was repeated for the other three capsules.

**Enrichment Culturing and Screening of Bacteria (Gram Negative vs. Gram Positive)**

About 350 ml of trypticase soy broth (TSB) was added to the cells recovered from each capsule filter and incubated at 37°C for 48 hours. After 48 hours of culture enrichment, the enrichment broth was divided into two equal aliquots with each aliquot poured into separate 500-ml Erlenmeyer flasks. One ml of phenylethyl alcohol (PEA) was added to one of the aliquots to select for Gram positive bacterial species while 0.1 ml of concentrated crystal violet solution (3.5g/liter) was added to the other aliquot to select for Gram negative bacterial species. PEA inhibits the growth of Gram negative bacteria by altering membrane permeability of cell and increasing leakage of cellular potassium (Silver and Wendt 1967, Leboffe and Pierce 2010). Crystal violet inhibits the growth of Gram positives by forming covalent adducts with thioredoxin reductase 2 (a highly conserved protein that is essential for cellular activity) once it penetrates the cell (Zhang et al. 2011). The flasks were shaken until a uniform mixture of culture and selective agent solution was achieved. Aluminum foil and parafilm were used to seal the flasks. This procedure was repeated for the other three samples from the other sites. All eight flasks were then incubated at 37°C for 48 hours.
**Pathogen Selection**

After 48 hours of incubation, each flask was shaken very well to suspend the bacterial cells in solution. Different media were employed to select for target pathogenic species as below and in Table 2.

**Table 2.** List of media used for enrichment and pathogen selection from four streams in Great Smoky Mountains National Park.

<table>
<thead>
<tr>
<th>Media</th>
<th>Classification</th>
<th>Type of Organism Used to Isolate</th>
</tr>
</thead>
<tbody>
<tr>
<td>Trypticase Soy Broth</td>
<td>Enrichment medium</td>
<td>All high nutrient requiring bacteria</td>
</tr>
<tr>
<td>Trypticase Soy Broth + Phenylethyl Alcohol</td>
<td>Selective medium</td>
<td>Gram positive bacteria</td>
</tr>
<tr>
<td>Trypticase Soy broth + Crystal Violet</td>
<td>Selective medium</td>
<td>Gram negative bacteria</td>
</tr>
<tr>
<td>Trypticase Soy Broth + 6.8% NaCl</td>
<td>Selective medium</td>
<td>Halotolerant bacteria</td>
</tr>
<tr>
<td>Mannitol Salt Agar (MSA)</td>
<td>Selective medium</td>
<td><em>Staphylococcus</em> species</td>
</tr>
<tr>
<td>Hektoen Enteric Agar (HE)</td>
<td>Selective medium</td>
<td><em>Salmonella</em> and <em>Shigella</em> species</td>
</tr>
<tr>
<td>MacConkey Agar</td>
<td>Selective medium</td>
<td>Gram negative enteric bacilli</td>
</tr>
<tr>
<td>Xylose Lysine Deoxycholate Agar (XLDA)</td>
<td>Selective medium</td>
<td><em>Salmonella, Shigella and Providencia</em> species</td>
</tr>
<tr>
<td>Salmonella Typhimurium Isolation Agar (STIA)</td>
<td>Selective medium</td>
<td><em>Salmonella typhimurium</em></td>
</tr>
<tr>
<td>Pseudomonas Isolation Agar (PIA)</td>
<td>Selective medium</td>
<td><em>Pseudomonas</em> species</td>
</tr>
</tbody>
</table>
Selection for Pseudomonas Species

*Pseudomonas* isolation agar (PIA) was prepared using dry premixed media and glycerol as indicated in the HIMEDIA product manual. One milliliter of Gram negative culture solution (TSB + 0.1 ml crystal violet solution + bacterial cells) was pipetted into a 2ml collection tube under sterile conditions. The collection tube was then placed on a high speed vortex to mix the cell suspension and a 100 μl aliquot was pipetted onto the solidified PIA media and spread-plated to uniformly cover the media with culture. This was done in triplicate and was repeated for the other three Gram negative culture solutions.

Selection for *Salmonella typhimurium*

*Salmonella typhimurium* isolation agar (STIA) was prepared according to the recipe laid out in the Handbook of Microbiological Media (Atlas 2010). The same protocol was used to spread Gram negative culture solution on the STIA plates as for the PIA media.

Selection for Gram negative Enteric Bacilli

MacConkey agar was used to screen for Gram negative enteric bacilli such as *Escherichia coli, Salmonella* species, *Shigella* species, and *Providencia* species. The same protocol utilized for PIA and STIA was employed for the MacConkey selection media. All media plates were incubated for 48 hours at 37°C. Unique bacterial colonies were selected and streaked onto three different media types: Hektoen enteric agar, xylose lysine tergitol-4 (XLT4) agar and xylose lysine deoxycholate agar. This procedure was done in triplicate and was repeated for the other three Gram negative culture solutions from all sites.
Selection for Staphylococcal Species

One hundred and twenty-five ml of Gram positive culture solution (TSB + PEA + bacterial cells) were added to 125ml of 13.6% NaCl TSB solution. This step yielded a 6.8% NaCl culture solution which was then incubated at 37°C for 48 hours. After incubation, a spread plate technique was utilized to culture bacterial colonies on mannitol salt agar plates. All plates were incubated for 48 hours at 37°C.
Figure 1: Schematic showing selective protocol sequence used to isolate potentially pathogenic bacterial species from water samples collected from four streams in Great Smoky Mountains National Park and the media type on which each isolate was isolated (key for isolates: NC =
North Carolina, T = Tennessee; KP = Kephart Prong, OR = Oconaluftee River, HKB = Hickory King Branch, LPR = Little Pigeon River; 1, 2, 3 = sequential isolates obtained from each site).

**Isolate Designations**

Isolates were initially labeled based on their location of origin and which media type they were isolated from, this made tracking each isolate easy but the names were too long. For simplification, each isolate was renamed based only on site of origin and culture number. For example, an isolate obtained from Oconaluftee River was labeled NCOR meaning North Carolina Oconaluftee River and a selection number, e.g. NCOR1.

**DNA Sequencing and Culture Matching**

Streak plating was employed to obtain pure cultures from different colonies with ten unique pure culture plates selected based on location of sample origin and growth on specific media types. These ten cultures (NCOR1, NCOR2, NCOR3, THKB1, THKB2, NCKP1, NCKP2, TLPR1, TLPR2 and TLPR3) were streaked onto separate trypticase soy agar plates and shipped to Genewiz Labs in South Plainville, New Jersey to have their 16S ribosomal DNA sequenced. Partial forward and reverse sequences were obtained for the 16S rRNA gene allowing for a nearly complete sequence to be assembled. Each sequence was assessed for quality using Finch TV version 1.4.0 (Geospiza, Inc.). Any base with a single distinct peak was defined as acceptable while those with multiple peaks of different colors were classified as unacceptable, if the peaks were of the same size. Where there was overlap between the two sequences, the corresponding DNA strand was assessed at the same position to determine the correct base for that position. The forward and reverse sequences were aligned using the National Center for Biotechnology Information (NCBI) Align Two Sequences Nucleotide BLAST tool (Altschul 1997). This alignment tool allowed for sequence proofreading of the overlapping areas of the
forward and reverse sequencing reads when paired with Finch TV’s graphical analysis tool. Corrected forward and reverse sequencing data were then combined into one sequence with approximately 1400 nucleotide bases.

The DNA sequence for each isolate was compared against sequences in the Ribosomal Database Project (RDP) (Cole et al. 2009). DNA sequencing data were first run through the Classifier tool (Wang et al. 2007) on the RDP website to obtain the hierarchical classification of each isolate from phylum to genus. The Sequence Match tool (SeqMatch) was then employed to compare each isolate’s DNA sequence to known species DNA sequences (Cole et al. 2014). To achieve this, the following parameters were set:

1. Strain option was limited to type strain only.
2. Only sequences from isolated cultures (not environmental DNA or RNA) were compared.
3. Only full length or close to full length (>1200 bases) sequences were selected to be compared from the database.
4. Sequence quality was set to good (excluded sequences of suspect or poor quality) and Taxonomy was set to nomenclatural (Cole et al. 2014).

This procedure produced a list of type bacterial species whose sequences were compared to each GSMNP isolate’s 16 rDNA sequence so that any possible relatives to the isolate could be determined.

Phylogenetic Analysis

To perform phylogenetic analysis, Seqmatch was utilized but one parameter was changed (strain parameter was changed from only type to both type and non-type) because the results of the initial search were limited to a few species. DNA sequences from the result of this search that
had a percent match >95% were used to generate a phylogenetic tree with the aid of the MUSCLE multiple sequence alignment tool (Edgar 2004), Gblocks conserved block selective tool (Castresana 2000) and PhyML phylogeny tool (Guindon et al. 2010). Gblocks eliminated poorly aligned positions and divergent regions in the aligned 16S rDNA sequences (Castresana 2000). Only well aligned positions were used to determine phylogeny. Bootstrapping procedure was set to 100 bootstraps and the GTR substitution model was selected and rendered via TreeDyn tool (Dereeper et al. 2008, Timothy Driscoll personal communication). The phylogenetic tree was constructed to aid in the identification of each isolate and to determine whether it was related to a pathogenic species.
RESULTS

Classification of Isolates

The RDP Classifier tool was able to categorize the isolates into specific genera with 100% confidence except for isolates TLPR2 and TLPR3 which had 83% and 81% percent confidence, respectively (Table 3). Isolates THKB1, NCKP2 and THKB2 were found to belong to the genus *Bacillus*, isolates NCOR2 and TLPR1 belonged to the genus *Proteus*, TLPR2 and TLPR3 belonged to the genus *Enterobacter*, NCOR1, NCKP1 and NCOR3 were members of the genera *Staphylococcus, Serratia* and *Providencia*, respectively.

**Table 3:** RDP Classifier results showing the genus each isolate obtained from GSMNP belonged to, and the accuracy (confidence) of the result in percentage.

<table>
<thead>
<tr>
<th>Isolate</th>
<th>Genus</th>
<th>% Confidence</th>
</tr>
</thead>
<tbody>
<tr>
<td>NCKP1</td>
<td><em>Serratia</em></td>
<td>100</td>
</tr>
<tr>
<td>NCKP2</td>
<td><em>Bacillus</em></td>
<td>100</td>
</tr>
<tr>
<td>NCOR1</td>
<td><em>Staphylococcus</em></td>
<td>100</td>
</tr>
<tr>
<td>NCOR2</td>
<td><em>Proteus</em></td>
<td>100</td>
</tr>
<tr>
<td>NCOR3</td>
<td><em>Providencia</em></td>
<td>100</td>
</tr>
<tr>
<td>THKB1</td>
<td><em>Bacillus</em></td>
<td>100</td>
</tr>
<tr>
<td>THKB2</td>
<td><em>Bacillus</em></td>
<td>100</td>
</tr>
<tr>
<td>TLPR1</td>
<td><em>Proteus</em></td>
<td>100</td>
</tr>
<tr>
<td>TLPR2</td>
<td><em>Enterobacter</em></td>
<td>83</td>
</tr>
<tr>
<td>TLPR3</td>
<td><em>Enterobacter</em></td>
<td>81</td>
</tr>
</tbody>
</table>

SeqMatch Results & Phylogenetic Trees

**NCKP1**

The RDP SeqMatch tool revealed that the isolate NCKP1’s 16S ribosomal DNA (rDNA) sequence had a 98.4% identity to a type strain of *Serratia marcescens* (KRED; accession number
AB061685) and 99.6% identity to a non-type strain *Serratia* sp. NT3. Phylogenetic analysis of NCKP1 (Figure 2a & b) showed *Serratia* sp. NT3 shares a common ancestor with the ancestor of isolate NCKP1.

**Table 4**: List of type strain bacterial species with ≥ 90% 16S rDNA sequence match to isolate NCKP1 and their percent identity to NCKP1.

<table>
<thead>
<tr>
<th>Isolate</th>
<th>Best SeqMatch Type Strain</th>
<th>Accession Number</th>
<th>% Identity to Isolate</th>
</tr>
</thead>
<tbody>
<tr>
<td>NCKP1</td>
<td><em>Serratia marcescens; KRED</em></td>
<td>AB061685</td>
<td>98.4</td>
</tr>
<tr>
<td></td>
<td><em>Serratia nematodiphila</em></td>
<td>EU036987</td>
<td>98.3</td>
</tr>
<tr>
<td></td>
<td><em>Serratia marcescens</em></td>
<td>AJ233431</td>
<td>96.4</td>
</tr>
<tr>
<td></td>
<td><em>Serratia ureilytica</em></td>
<td>AJ854062</td>
<td>92.8</td>
</tr>
</tbody>
</table>

**Figure 2a**: Phylogenetic tree constructed using the top 12 sequence matches from the Ribosomal Database Project (type & non-type strains) analyzed with MUSCLE multiple sequence alignment tool, Gblocks selective tool, PhyML phylogeny tools and rendered via TreeDyn tool for isolate NCKP1 from Great Smoky Mountains National Park.
Figure 2b: Phylogenetic tree constructed using the top 12 sequence matches from the Ribosomal Database Project (type & non-type strains), and analyzed with MUSCLE multiple sequence alignment tool, Gblocks selective tool, PhyML phylogeny tool and rendered via TreeDyn tool with branch distance ignored for isolate NCKP1 from Great Smoky Mountains National Park. Branch support values are out of 1.0.

TLPR1 & NCOR2

The 16S rDNA sequence for isolate TLPR1 had a 98.1% identity to type strain *Proteus mirabilis* (accession number DQ885256) while isolate NCOR2 had a 97.6 % identity to the same type strain when its 16S rDNA sequence was run through SeqMatch. Phylogenetic analysis revealed TLPR1 was related to the ancestor of non-type strain *Proteus penneri* (YCY34) while NCOR2 was shared a common ancestor with non-type strains *Proteus mirabilis* (SPC04) and *Proteus mirabilis* (FUA 1240 5b1).
**Table 5:** List of type strain bacterial species with ≥ 90% 16S rDNA sequence match to isolates TLPR1 and NCOR2 and their percent identity to TLPR1 and NCOR2.

<table>
<thead>
<tr>
<th>Isolate</th>
<th>Best SeqMatch Type Strain Results</th>
<th>Accession Number</th>
<th>% Identity to Isolate</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>NCOR2</td>
<td><em>Proteus mirabilis</em></td>
<td>DQ885256</td>
<td>97.6</td>
<td></td>
</tr>
<tr>
<td></td>
<td><em>Proteus vulgaris</em></td>
<td>DQ885257</td>
<td>93.3</td>
<td></td>
</tr>
<tr>
<td></td>
<td><em>Proteus penneri</em></td>
<td>DQ885258</td>
<td>91.8</td>
<td></td>
</tr>
<tr>
<td></td>
<td><em>Proteus hauseri</em></td>
<td>DSM14437</td>
<td>91.1</td>
<td></td>
</tr>
<tr>
<td>TLPR1</td>
<td><em>Proteus mirabilis</em></td>
<td>DQ885256</td>
<td>98.1</td>
<td></td>
</tr>
<tr>
<td></td>
<td><em>Proteus vulgaris</em></td>
<td>DQ885257</td>
<td>93.8</td>
<td></td>
</tr>
<tr>
<td></td>
<td><em>Proteus penneri</em></td>
<td>DQ885258</td>
<td>92.1</td>
<td></td>
</tr>
<tr>
<td></td>
<td><em>Proteus hauseri</em></td>
<td>DSM14437</td>
<td>91.6</td>
<td></td>
</tr>
</tbody>
</table>

**Figure 3a:** Phylogenetic tree constructed using the top 9 sequence matches from the Ribosomal Database Project (type & non-type strains), and analyzed with MUSCLE multiple sequence alignment tool, Gblocks selective tool, PhyML phylogeny tool and rendered via TreeDyn tool for isolates NCOR2 and TLPR1 from Great Smoky Mountains National Park.
Figure 3b: Phylogenetic tree constructed using the top 9 sequence matches from the Ribosomal Database Project (type & non-type strains), and analyzed with MUSCLE multiple sequence alignment tool, Gblocks selective tool, PhyML phylogeny tool and rendered via TreeDyn tool with branch distance ignored for isolates NCOR2 & TLPR1 from Great Smoky Mountains National Park. Branch support values are out of 1.0.

**NCOR1**

The RDP SeqMatch tool revealed that the 16S rDNA sequence for isolate NCOR1 had a 99% identity to type strain *Staphylococcus sciuri* (accession number AJ421446) and a 99.8% identity to non-type strain *Staphylococcus* sp. K4-STE/2013. Phylogenetic analysis showed NCOR1 shares a common ancestor with the ancestor of type strain *Staphylococcus sciuri* (accession number AJ421446) and non-type strain *Staphylococcus sciuri* subsp. sciuri (P3-3-b-1).
**Figure 4a:** Phylogenetic tree constructed using the top 9 sequence matches from the Ribosomal Database Project (type & non-type strains), and analyzed with MUSCLE multiple sequence alignment tool, Gblocks selective tool, PhyML phylogeny tool and rendered via TreeDyn tool for isolate NCOR1 from Great Smoky Mountains National Park.

**Figure 4b:** Phylogenetic tree constructed using the top 9 sequence matches from the Ribosomal Database Project (type & non-type strains), and analyzed with MUSCLE multiple sequence alignment tool, Gblocks selective tool, PhyML phylogeny tool and rendered via TreeDyn tool with branch distance ignored for isolate NCOR1 from Great Smoky Mountains National Park. Branch support values are out of 1.0.
Table 6: List of type strain bacterial species with ≥ 90% 16S rDNA sequence match to isolate NCOR1 and their percent identity to NCOR1

<table>
<thead>
<tr>
<th>Isolate</th>
<th>Best SeqMatch Type Strain Results</th>
<th>Accession Number</th>
<th>% Identity to Isolate</th>
</tr>
</thead>
<tbody>
<tr>
<td>NCOR1</td>
<td>Staphylococcus sciuri</td>
<td>AJ421446</td>
<td>99</td>
</tr>
<tr>
<td></td>
<td>Staphylococcus sciuri</td>
<td>AB233331</td>
<td>98.2</td>
</tr>
<tr>
<td></td>
<td>Staphylococcus sciuri</td>
<td>AB233332</td>
<td>96.9</td>
</tr>
<tr>
<td></td>
<td>Staphylococcus lentus</td>
<td>D83370</td>
<td>94.6</td>
</tr>
<tr>
<td></td>
<td>Staphylococcus vitulinus</td>
<td>AB009946</td>
<td>93.4</td>
</tr>
</tbody>
</table>

NCOR3

The 16S rDNA sequence for isolate NCOR3 had a 98.3% identity to type strain Providencia rettgeri (accession number AM040492). Phylogenetic analysis revealed NCOR3 shares a common ancestor with non-type strain Providencia rettgeri VITJCSTT3.

Table 7: List of type strain bacterial species with ≥ 90% 16S rDNA sequence match to isolate NCOR3 and their percent identity to NCOR3

<table>
<thead>
<tr>
<th>Isolate</th>
<th>Best SeqMatch Type Strain Results</th>
<th>Accession Number</th>
<th>% Identity to Isolate</th>
</tr>
</thead>
<tbody>
<tr>
<td>NCOR3</td>
<td>Providencia rettgeri</td>
<td>AM040492</td>
<td>98.3</td>
</tr>
<tr>
<td></td>
<td>Providencia vermicola</td>
<td>AM040495</td>
<td>96.8</td>
</tr>
<tr>
<td></td>
<td>Providencia burhodogranariae</td>
<td>HM038004</td>
<td>94</td>
</tr>
<tr>
<td></td>
<td>Providencia rustigianii</td>
<td>AM040489</td>
<td>93.1</td>
</tr>
<tr>
<td></td>
<td>Providencia sneebia</td>
<td>HM038003</td>
<td>92.5</td>
</tr>
</tbody>
</table>
**Figure 5a:** Phylogenetic tree constructed using the top 8 sequence matches from the Ribosomal Database Project (type & non-type strains), and analyzed with MUSCLE multiple sequence alignment tool, Gblocks selective tool, PhyML phylogeny tool and rendered via TreeDyn tool for isolate NCOR3 from Great Smoky Mountains National Park.

**Figure 5b:** Phylogenetic tree constructed using the top 8 sequence matches from the Ribosomal Database Project (type & non-type strains), and analyzed with MUSCLE multiple sequence alignment tool, Gblocks selective tool, PhyML phylogeny tool and rendered via TreeDyn tool for isolate NCOR3 from Great Smoky Mountains National Park. Branch support values are out of 1.0.
Both isolates THKB2 and NCKP2 16S rDNA sequences had a high percent identity to type strain *Bacillus subtilis* (accession number AJ276351). THKB2 had 99.4% identity to this bacteria strain while NCKP2 had a 99.5% identity with the strain. Phylogenetic analysis of both 16S rDNA sequences revealed THKB2 and NCKP2 share a common ancestor. This ancestor shares a common ancestor with *Bacillus subtilis* (accession number AJ276351).

<table>
<thead>
<tr>
<th>Isolate</th>
<th>Best SeqMatch Type Strain Results</th>
<th>Accession Number</th>
<th>% identity to Isolate</th>
</tr>
</thead>
<tbody>
<tr>
<td>NCKP2</td>
<td><em>Bacillus subtilis</em></td>
<td>AJ276351</td>
<td>99.5</td>
</tr>
<tr>
<td></td>
<td><em>Bacillus mojavensis</em></td>
<td>AB021191</td>
<td>98</td>
</tr>
<tr>
<td></td>
<td><em>Bacillus vallismortis</em></td>
<td>AB021198</td>
<td>97.4</td>
</tr>
<tr>
<td></td>
<td><em>Bacillus subtilis</em></td>
<td>AF074970</td>
<td>97.3</td>
</tr>
<tr>
<td></td>
<td><em>Bacillus methylophilus</em></td>
<td>EU194897</td>
<td>97.1</td>
</tr>
<tr>
<td>THKB2</td>
<td><em>Bacillus subtilis</em></td>
<td>AJ276351</td>
<td>99.4</td>
</tr>
<tr>
<td></td>
<td><em>Bacillus mojavensis</em></td>
<td>AB021191</td>
<td>98.1</td>
</tr>
<tr>
<td></td>
<td><em>Bacillus subtilis</em></td>
<td>AF074970</td>
<td>97.7</td>
</tr>
<tr>
<td></td>
<td><em>Bacillus vallismortis</em></td>
<td>AB021198</td>
<td>97.2</td>
</tr>
</tbody>
</table>
**Figure 6a:** Phylogenetic tree constructed using the top 8 sequence matches from the Ribosomal Database Project (type & non-type strains), and analyzed with MUSCLE multiple sequence alignment tool, Gblocks selective tool, PhyML phylogeny tool and rendered via TreeDyn tool for isolates THKB2 & NCKP2 from Great Smoky Mountains National Park.

**Figure 6b:** Phylogenetic tree constructed using the top 8 sequence matches from the Ribosomal Database Project (type & non-type strains), and analyzed with MUSCLE multiple sequence alignment tool, Gblocks selective tool, PhyML phylogeny tool and rendered via TreeDyn tool with branch distance ignored for isolates THKB2 & NCKP2 from Great Smoky Mountains National Park. Branch support values are out of 1.0.
RDP SeqMatch search results revealed isolate THKB1 16S rDNA sequence had a 99.3% identity to type strain *Bacillus licheniformis* (accession number CP000002). THKB1 and non-type strain *Bacillus* sp. HY13 2010 share a common ancestor according to phylogenetic analysis.

**Table 9:** List of type strain bacterial species with ≥ 90% 16S rDNA sequence match to isolate THKB1 and their percent identity to THKB1.

<table>
<thead>
<tr>
<th>Isolate</th>
<th>Best SeqMatch Type Strain Results</th>
<th>Accession Number</th>
<th>% identity to Isolate</th>
</tr>
</thead>
<tbody>
<tr>
<td>THKB1</td>
<td><em>Bacillus licheniformis</em></td>
<td>CP0000002</td>
<td>99.3</td>
</tr>
<tr>
<td></td>
<td><em>Bacillus aerius</em></td>
<td>AJ831843</td>
<td>97.2</td>
</tr>
<tr>
<td></td>
<td><em>Bacillus sonorensis</em></td>
<td>AF302118</td>
<td>95.4</td>
</tr>
</tbody>
</table>

**Figure 7a:** Phylogenetic tree constructed using the top 8 sequence matches from the Ribosomal Database Project (type & non-type strains), and analyzed with MUSCLE multiple sequence alignment tool, Gblocks selective tool, PhyML phylogeny tool and rendered via TreeDyn tool for isolate THKB1 from Great Smoky Mountains National Park.
Figure 7b: Phylogenetic tree constructed using the top 8 sequence matches from the Ribosomal Database Project (type & non-type strains), and analyzed with MUSCLE multiple sequence alignment tool, Gblocks selective tool, PhyML phylogeny tool and rendered via TreeDyn tool with branch distance ignored for isolate THKB1 from Great Smoky Mountains National Park. Branch support values are out of 1.0.

**TLPR2 & TLPR3**

Isolates TLPR2 and TLPR3 had 16S rDNA sequences with 95.9 % and 95.6% identity, respectively, to *Enterobacter hormaechei* (accession number AJ508302). Phylogenetic analysis of both isolates showed TLPR2 shares a common ancestor with non-type strain *Enterobacter cloacae* (U26978) while TLPR3 shares a common ancestor with *Enterobacter ludwigii* (EN-119).
**Table 10:** List of type strain bacterial species with ≥ 90% 16S rDNA sequence match to isolates TLPR2 and TLPR3 and their percent identity to TLPR2 and TLPR3.

<table>
<thead>
<tr>
<th>Isolate</th>
<th>Best SeqMatch Type Strain Results</th>
<th>Accession Number</th>
<th>% identity to Isolate</th>
</tr>
</thead>
<tbody>
<tr>
<td>TLPR1</td>
<td><strong>Enterobacter hormaechei</strong></td>
<td>AJ508302</td>
<td>95.9</td>
</tr>
<tr>
<td></td>
<td><strong>Enterobacter ludwigii</strong></td>
<td>AJ853891</td>
<td>94.5</td>
</tr>
<tr>
<td></td>
<td><strong>Enterobacter cancerogenus</strong></td>
<td>LMG 2693</td>
<td>93.9</td>
</tr>
<tr>
<td></td>
<td><strong>Enterobacter asburiae</strong></td>
<td>AB004744</td>
<td>93.8</td>
</tr>
<tr>
<td>TLPR3</td>
<td><strong>Enterobacter hormaechei</strong></td>
<td>AJ508302</td>
<td>95.6</td>
</tr>
<tr>
<td></td>
<td><strong>Enterobacter ludwigii</strong></td>
<td>AJ853891</td>
<td>93.1</td>
</tr>
<tr>
<td></td>
<td><strong>Enterobacter cancerogenus</strong></td>
<td>LMG 2693</td>
<td>92.6</td>
</tr>
<tr>
<td></td>
<td><strong>Enterobacter asburiae</strong></td>
<td>AB004744</td>
<td>92.4</td>
</tr>
<tr>
<td></td>
<td><strong>Pantoea septica</strong></td>
<td>EU216734</td>
<td>92.1</td>
</tr>
</tbody>
</table>

*Enterobacter ludwigii KS81

**Enterobacter ludwigii** (T); type strain; EN-119

**Figure 8a:** Phylogenetic tree constructed using the top 8 sequence matches from the Ribosomal Database Project (type & non-type strains), and analyzed with MUSCLE multiple sequence alignment tool, Gblocks selective tool, PhyML phylogeny tool and rendered via TreeDyn tool for isolates TLPR2 & TLPR3 from Great Smoky Mountains National Park.
* Enterobacter ludwigi KS81

** Enterobacter ludwigi (T); type strain; EN-119

** Figure 8b: Phylogenetic tree constructed using the top 8 sequence matches from the Ribosomal Database Project (type & non-type strains), and analyzed with MUSCLE multiple sequence alignment tool, Gblocks selective tool, PhyML phylogeny tool and rendered via TreeDyn tool with branch distance ignored for isolates TLPR2 & TLPR3 from Great Smoky Mountains National Park. Branch support values are out of 1.0.
DISCUSSION

Isolate Selection and Classification

The results suggest that initial enrichment and screening of bacteria via phenylethyl alcohol and crystal violet was successful. Isolates that were obtained through the Gram negative pipeline were related at the genus and species level to known Gram negative bacteria. Isolates obtained using the Gram positive pipeline were related to Gram positive bacterial species with the exception of THKB1, which was related to Gram positive bacteria but grew on Pseudomonas isolation agar. This was unexpected but recent literature on triclosan (selective agent in PIA media) revealed that many bacterial species including Staphylococcus aureus (a Gram positive species) and Salmonella enterica have developed resistance to triclosan because it is used as an antiseptic agent in hand soap and even in toothpaste (Carey & McNamara 2014). NCOR1, THKB2 and NCKP2 were related to Gram positive bacterial species while NCOR2, NCOR3, NCKP1, TLPR1, TLPR2 and TLPR3 were related to Gram negative bacterial species. Even though a limited number of isolates were selected and sequenced; sequencing results suggest a relatively high amount of species diversity considering the selective media that were employed.

According to the RDP Classifier tool results, all the Gram negative isolates belonged to the phylum Proteobacteria and family Enterobacteriaceae while the Gram positive isolates were in the phylum Firmicutes but are members of two different families, Staphylococcaceae and Bacillaceae. THKB1, THKB2 and NCKP2 are members of family Bacillaceae while NCOR1 belonged to family Staphylococcaceae. This result was expected because all the selective media used in this research to select for the isolates are designed to culture known pathogenic bacteria belonging to these families.
Isolate NCKP1 was initially identified as *Salmonella typhimurium* because it was the only bacterium that grew on the STIA plates. It was a surprise when analysis of the NCKP1 16S rDNA sequence using both a phylogenetic tree and RDP tools suggested NCKP1 is a member of the genus *Serratia* (Table 3 & Figure 2a & b).

*Serratia* spp. are rod shaped Gram negative bacteria and some members of this genus produce pigments named prodigiosins (Hearn et al. 1970; Gerber 1975). Certain species of *Serratia* such as *S. marcescens* have clinical importance and this species is known to cause nosocomial infections (Grimont & Grimont 1992; Brenner 1984). *Serratia marcescens* are opportunistic pathogens that are able to form biofilms and cause respiratory and urinary tract infections (Yu 1979). SeqMatch analysis of NCKP1 matched the isolate with *S. marcescens* KRED; a type strain subspecies that was isolated from a wastewater treatment tank in Saku, Japan (Ajithkumar et al. 2003).

Even though isolate NCKP1 had a 98.4% identity to *S. marcescens* KRED, it did not exhibit the same red-colored pigment (prodigiosin) described in 2003 by Ajithkumar et al. about *S. marcescens* KRED cultured on trypticase soy agar (TSA) plates at room temperature. The reason isolate NCKP1 was cultured at 25°C was because *S. marcescens* strains are known to lose their pigmentation at higher temperatures (Rjazantseva et al. 1994). Phylogenetic analysis showed isolate NCKP1 to be a branched descendent of *S. marcescens* KRED and non-type strain *Serratia* sp. NT3. The isolate is also related to the ancestor of three *S. marcescens* subspecies (Figure 2b). Molecular genetic analysis of isolate NCKP1 revealed the isolated was a member of the genus *Serratia*, and was related to *S. marcescens*. Metabolic testing and environmental
parameter susceptibility testing is required to confirm how closely NCKP1 is related to *S. marcescens*.

**TLPR1 & NCOR2**

Both isolates TLPR1 and NCOR2 were isolated on Hektoen enteric agar plates, and 16S rDNA sequences matched these two isolates to the same genus and type strain species. Yet maximum likelihood phylogenetic analysis placed both isolates at different positions on the tree and TLPR1 was revealed to be the ancestor of NCOR2 (Figure 3a & b). Isolate TLPR1 and NCOR2 belong to the genus *Proteus*; and are related to *P. penneri* and *P. mirabilis*, respectively, according to phylogenetic analysis.

RDP SeqMatch on the other hand matched both isolates to type strain *Proteus mirabilis* (NCTC 11938). *Proteus mirabilis* and *P. penneri* are Gram negative facultative anaerobic bacilli that are known to cause nosocomial urinary infections in immunocompromised patients (O'Hara et al. 2000). These *Proteus* species are closely related and are sometimes misidentified (Kishore 2012). *Proteus mirabilis* strains are much more prevalent in soils and freshwater bodies than *P. penneri* and cause more infections (Cantón et al. 2006).

**NCOR1**

The genus *Staphylococcus* has more than 40 species and subspecies, with a majority of the species being opportunistic pathogens (Marsou et al. 1999). Staphylococci are widespread in nature and are known to cause infections in both humans and animals (Kloos & Schleifer 1975; Kloos & Bannerman 1995). Perhaps the most notable species of this genus is *S. aureus* which can cause methicillin-resistant infections commonly known as MRSA (Kloos & Bannerman 1995).
*Staphylococcus sciuri* was initially classified as an animal pathogen that caused infections in livestock and rodents as seen in the subspecies *S. sciuri* subsp. *rodentium* (Kloos & Schleifer 1975). In recent years, *S. sciuri* has crossed over to humans and is now observed to cause infections in humans (Kalawole & Shittu 1997). Based on the research findings in 1999 by Marsou et al., *S. sciuri* is a natural reservoir of the *mecA* gene. This gene is responsible for the methicillin resistance observed in *S. aureus* (Wu et al. 1996; Vandenesch et al. 2003). Results suggest isolate NCOR1 is a member of genus *Staphylococcus* and has a 99% identity to *S. sciuri*. This means NCOR1 could potentially be pathogenic to humans.

**NCOR3**

This isolate was obtained via the Gram negative bacteria selection pipeline and was isolated on *Pseudomonas* isolation agar (PIA). Unlike isolate NCKP1, it was not a surprise when 16S rDNA sequencing analysis revealed the isolate belonged to a different genus – *Providencia*, instead of *Pseudomonas* – because the isolate did not exhibit the distinct green to yellow-green or sometimes pink coloration of colonies that is observed in true *Pseudomonas* spp. when cultured on PIA plates (S. O’Connell, personal communication).

*Providencia rettgeri* is a ubiquitous opportunistic pathogen that causes disease in a variety of animals including humans (Ladds et al. 1996 and Traub et al. 1971). In humans however, *P. rettgeri* has mainly been associated with causing nosocomial urinary tract infections (UTIs), but can also cause diarrhea and ocular infections such as keratitis, although this is rare (Koreishi et al. 2006). Most *P. rettgeri* strains are susceptible to antibiotic and antimicrobial agents, but there have been cases of resistance observed in the last decade (Chander et al. 2006). On the phylogenetic tree constructed, NCOR3 shares a direct common ancestor with a non-type
strain of *P. rettgeri* (Figure 5a & b). NCOR3 could be a strain or subspecies of *P. rettgeri* given that it has a 98.3% identity to the species and is closely related to one of its strains.

**THKB2, NCKP2 & THKB1**

Three isolates were found to be related to the genus *Bacillus* in this study. *Bacilli* are ubiquitous and can be found abundantly in soil and water bodies. THKB2 and NCKP2 were isolated on MSA plates while isolate THKB1 was cultured on a PIA plate. THKB2 and NCKP2 had rapid observable growth rates compared to the other isolates and formed lawns on the MSA plates. Neither isolate fermented mannitol as the plates remained red and no color change was observed. On PIA plates isolate THKB1 presented a creamy yellow pigmentation. 16S rDNA analysis matched THKB2 and NCKP2 to *Bacillus subtilis* (>99% similarity) while THKB1 showed the highest similarity to *Bacillus licheniformis* (>99%). On phylogenetic trees (Figure 6a & 6b), THKB2 and NCKP2 share a common ancestor; which implies these two isolates are closely related. Dissimilar to THKB2 and NCKP2, THKB1 was further away from its highest (*Bacillus licheniformis*) match on the phylogenetic tree. It shared a common ancestor with *Bacillus* sp. HY13 2010 (>99% similarity).

*B. subtilis* are benign bacteria that form part of the natural microflora of the gastrointestinal tract in humans. In Italy, *B. subtilis* is commonly utilized as a probiotic to treat and/or prevent intestinal disorders, and can be purchased at local pharmacies (Mazza et al. 1992 and Oggioni et al. 1998). It is not recognized as a pathogen, but there was a rare case presented in 1998 by Oggioni et al. about a 73 year old man with chronic lymphocytic leukemia who developed a recurring case of *B. subtilis* induced septicemia. *B. licheniformis* on the other hand is an opportunistic pathogen that has been associated with food poisoning in humans and bovine toxemia and abortions (Salkinoja-Salonen et al. 1999 and Logan 2012).
**TLPR2 & TLPR3**

As previously mentioned in the results section, the RDP Classifier tool was unable to determine the genus that isolates TLPR2 and TLPR3 belonged to with 100% confidence. The isolates were classified as *Enterobacter* at 83% and 81% confidence, respectively. Many species that belong to the family Enterobacteriaceae are difficult to identify via 16S rDNA alone because of the massive diversity within the family and diverse phenotypic characteristics (S. O’Connell, personal communication). Similar to *Staphylococcus*, most members of the genus *Enterobacter* are thought to be potentially pathogenic and cause opportunistic community and nosocomial infections (Sanders and Sanders 1997).

Although both isolates showed high similarity to *Enterobacter hormaechei* (>95%), phylogenetic analysis of isolates TLPR2 and TLPR3 revealed these two isolates were related to different species. TLPR2 shared a common ancestor with a non-type strain of *Enterobacter cloacae* while TLPR3 shared a common ancestor with *Enterobacter ludwigii*. *Enterobacter hormaechei* and *E. cloacae* are members of the *Enterobacter cloacae* complex (a group of six genetically and phenotypically similar *Enterobacter* species); and in 2005, Hoffmann et al. proposed a novel species, *E. ludwigii*, to join the complex (Hoffmann and Roggenkamp 2003). All three species of *Enterobacter* are opportunistic pathogens that cause nosocomial infections in immunocompromised patients (Mayhall et al. 1979 and Hoffmann et al. 2005).
CONCLUSION

A positive confirmation that an isolate was a potential pathogen was defined by the following: An isolate that had a 100% SeqMatch identity with a well-characterized pathogen, based on 16S rDNA sequence data, and shared an immediate ancestor with that pathogen on a phylogenetic tree would be considered as a potential pathogen as it was highly related to that species. The results revealed that none of the isolates sequenced had a 100% match to an obligate or opportunistic pathogen but a few isolates including NCOR3, TLPR2 and TLPR3 shared an immediate common ancestor with the opportunistic pathogens *P. rettgeri, E. cloacae* and *E. ludwigi*, respectively based on phylogeny. Most of the isolates with the exception of THKB2 and NCKP2 were most closely related to opportunistic pathogens.

Although none of the isolates had a 100% match to known pathogenic species, many of them matched (>97%) to opportunistic pathogens, which generally supports the hypothesis that animal pathogens might be found in this study. By general consensus, isolates with high sequencing match (≥99%) to a well documented bacterial species are considered to be closely related or sometimes identified as that bacterial species while a 97 to 99% identity in 16S rDNA is a criterion used to identify an organism at genus level and sometimes species level (Drancourt et al. 2004). Since many of the isolates fell between 97 to 99% identity, further tests are needed to confirm the true identity and pathogenicity of each species.

Some tests that could be used to determine bacterial pathogenicity and infectivity include virulence gene cloning (Wu et al. 1996), running a susceptibility test to both natural and synthetic antibiotics and antimicrobials (as described in Edberg et al. 1996), or whole genome sequencing and probing for virulence genes. Infectivity could be tested simply by exposing
healthy mice to the isolate either via ingestion or injection of the isolate into the blood stream of the mice. Virulence factors such as exotoxins could also be tested via immunoassays such as enzyme-linked immunosorbent assay (ELISA) to for specific exotoxins (Kato et al. 1998, Zhu et al. 2014).

This study supports the idea that pathogenic bacteria exist in Great Smoky Mountains National Park. There were many isolates obtained during the selection process but due to time and resource constraints only 10 isolates were selected, thus there are pathogens that could have been missed in this work. Without these constraints, more isolates could have been cultured from the initial screening media (TSB + PEA and TSB + crystal violet). Isolates obtained would have undergone both 16S sequencing and phylogenetic analysis. And once the genus for each isolate was identified, the isolates would undergo a series of metabolic and environmental susceptibility test to confirm their identity. Pathogenicity would be tested using one or more of the methods mentioned above. Conversely a fully molecular approach could be used to identify bacterial pathogens that could be in the streams. This approach would rely on polymerase chain reaction (PCR) and microarray based assay as described in Järvinen et al. 2009.

This study used methods associated with culturing human pathogens. For this reason, many animal pathogens might have been selected against. This would be especially so for non-mammalian and avian species. Different selection procedures and variables could be employed such as the use of different selective media types and incubation temperature could be changed to find pathogens that were selected against using the current procedures.

It is possible that the four sites sampled for pathogens may not be reservoirs for pathogens or are not conducive for pathogen proliferation. Other sites such as camping grounds and stagnant water bodies in the park could be better sites for isolating human pathogens. This is
because the continuous flow of water could prevent the pathogens from settling in and proliferating while stagnated water-bodies are nutrient rich and conducive for growth; camping grounds are also more likely to harbor pathogens since there is a higher probability that people travel to those sites may unknowingly be carrying pathogens on their skin, clothes and could also introduce microbes from their food sources and untreated wastes.
REFERENCES


APPENDIX

Water Pump Description:
The water pump used to this study was a hand drill powered pump (Flotec FPDMP21SA-P2 self priming drill pump) with about 5 ft of plastic hose attached to one end and a Pall Gelman Sciences capsule filter clamped to the other end. The free end of the hose had a steel mesh screwed to it. The hose and steel mesh were attached to two large plastic colanders of equal size (~ 8 inches in diameter) clamped together to form a sphere. A different hose was attached to the outlet end of the capsule filter. This hose emptied into a 6-gallon jerry can for measuring the volume of water sampled.
16S rDNA Sequences from ten bacterial isolates obtained from streams in Great Smoky Mountains National Park (key for isolates: NC = North Carolina, T = Tennessee; KP = Kephart Prong, OR = Oconaluftee River, HKB = Hickory King Branch, LPR = Little Pigeon River; 1, 2, 3 = sequential isolates obtained from each site).

NCOR1

1 CTATACATGC AGTCGAGCGA ACAGATGAGA AGCTTGCTTC TCTGATGTTA GCAGCGGACG
61 GGTGAGTAAAC ACCTGCTGTT ACACCTCAG AACTCCGGGA AACCAGGGCT
121 AATACCGGAT AATATTGCAG ACCCATCGGT TCAATAGTG AAGACGTTTG CCGCTGTAC
181 TTATAGATGG ACCCGGCCTG TATTAGCTAG TTGGTAAAGT AACGGCTTAC CAAGGCGACG
241 ATACGTAGGCC GACCTGAGAG GGTGATCGGC CACACTGGAA CTGAGCACGC GTCCAGACTC
301 CTTACGGGAGG CAGCAGATGG TGGCAAGCGT TATCCGGGA A TTATTGGGCG TAAAGCGCGC
361 GCGTGAGTGTGA GAGGTCTTT GGGCTGTCAC TTATAGATGG ACCCGGCCTG TATTAGCTAG TTGGTAAAGT AACGGCTTAC CAAGGCGACG
421 TAGTAACTGACAAGTCTTGG ACCGTACACT ACCAGAAAGC CACGGACTAC TACGTGCGAG
481 CGCCCGCGAT ATACGTAGG TGGCAAGCGT TATCCGGGA A TTATTGGGCG TAAAGCGCGC
541 GTAGGGCGGT TCTTAACTTCTG ATGTGAAAC CTCCACGGCTC AACCCTGAGG GGTCATTGGA
601 AACTGGGAAA CTTGAGTCAA GAAAGACAG GGTGAATTCC ATGTTAGGCC GTGAAATGCG
661 CAGAGATAGT GAGGAACACC AGTGGCCAAG GCCGCTCTCT GTGCTGATAC TGACAGCTGAT
721 GTGCGAAAAGC GTGGGAGATCA AACAGGATTA GATACCTTGG TAGTCCACGC GTGAAACGAT
781 GAGTTGCTAAAG TGTAGGCGGG TTTCCGCCCC TTATGCTGCA AGCATACGC A TAAACGACT
841 CGCCCTGGGGA GTACGACCGG CAGGTTGAAA CTCAAGGAA TTGACGGGGA CCGCCAAG
901 CGGTGGAGCA TGTTGGTTAA TTCGCAAGCAA CGCGAGAAAC CTGACCATCC CTTAAGGAAA
961 TTTGACCGCT CTAGAGATAG TGGCAAGCGT TATCCGGGA A TTATTGGGCG TAAAGCGCGC
1021 GTGTGCTGCA GCTGCTGTGCG TAGATTTGTT GTTTAAGTCC CGCAACGAGC GCAACCCCTA
1081 AGCTTATTTG CCACTATTAA GTTGGGGACT CTAGGTGACT TGGCGGGAAG AAACCGAGG
1141 AAGGTGGGAA TGGACCTCAA CTACATGCC CTTATGATT TGGGCTTAC ACCTGCTACA
1201 ATGGATAATA CAAGGACCAG CAGAATCCCGC AGGCCAACGA AATCCCATAA AATTATTCTC
1261 AGTCTGGATT GATAGTCTGCA ACTCGACTAC ATGAAGCTGG AATCGCTAGT AATCGCTAGT
1321 CAGCATCGCTA CGGTGATCAC GTTCCCCGGGT CTTGTAACA CCAGCGTCA CACACGGGAC
1381 GTTTGTAACA CCCGAAGCGG GTTGGAGTAAAC CTATAGGAGC TAGC

43
NCOR2
1 TGCAGTCGAG CGGTAAACAGG AGAAAGCTTG CTTTCTTGCT GACGAGCGGC GGACGGGTGA
61 GTAATGTATG GGGATCTGCC CGATAGAGGG GGATAACTA C TGGAAACGGT GGCTAATACC
121 ACATAATGTC TACGGACCAA AGCAAGGCGCT CTTCGGACCT TGCACTATCG GATGAAACCA
181 TAATGGATTT GCTAGTATGGG GGGTAAAGGG CTCACCTAGG CGACGATCTC TAGCTGGTCT
241 GAGAGGATGAC TACGCCCAAC TGGGACTGAG ACACGCGCCA GACTCTACGG GAGAGGAGCA
301 GTGGGGAATA TTGCAACAATG GGCACAAGCC TGATGCAGGC ATGCCCAGTG TATGAGAAG
361 GCCTTAGGGT TGTAAGTATC TTTCAGCGGG GAGGAAAGTG AGGAAGTTAA TACCCTTATC
421 ACCTGGACTTT ACCGCGCAAG GAAGACCGGG CTAACTCCGG GCCAGCAGGC GCCGTAATAC
481 GGAGGGTGCA AGGCTTAATC GGAATTACTG GGCCTAAAGG GCACGCAGGC GGTCAATTAA
541 GTGAGATGTG AAAGCCCGGA GCTCAATCTTG CTAAGAGCAT CTAAGAGCAT TTGGCTAGAG
601 TCTGTAGAG GGGGTAGAAA TTCCATGTGT AGCGGTGAAA TCGTATAGA GGTGAGGAAG
661 TACCGGTGCC GAAGCGGCGCC CCCTGGACAA AGACTGACGC TCAGGTCGGA AAGCGTGGCC
721 AGCAAAACAGG ATTAGATACC CTGGTAGTCC ACGGCTGAAAA CGATGTCGAT TTGAGGTTG
781 TGGTCCTGAA CGGTGCTTTC TGGAGCTAAC GCGTTAAATC GACGGCCTGG GAGTACGGGC
841 CGCAAGGTATA AAACCTCAAT GAATTGACGG GGGCCGCGAC AAGCGGTGGA GCAATGTTGTT
901 TAATTCGATG CAACCGCAGG AACCTCAGAC ACTCTCAGAC TCCAGCGAAA ATCTTCTGGA
961 TAGAGGAGTT CCTCGGGGAA CGCAGGCTGA CCTCGGTGCA ATGCCTGCGT TTGAGGAGAT
1021 GTGAAATGTT GGGTAAAGTC CGCAAGCAGG CGCAACCTCT ATCTTTCTGG GCCAGCAGGT
1081 AATGGTGGGA ACTCAAGAGA GACTGCGGGT GATAAACCAG AGGAAGGTGG GATGACGTCC
1141 AAGTCACTCAT GGGCCCTTAC AGTAGGGGCT AACACGTGCT ACAATGCGAG AATCAAAAGAG
1201 AAGCGACCTG GCAGAGCGAA GCAGAACTCA TAAAGTCTGT CGTAGTCGCG ATTGGAGTCT
1261 GCAACTCAGG TCCATGAAGT GGAATCGGT AGTAATCGTA GATCAGAGATG CTACGGTGAA
1321 TACGTTCGGG GGGCTTTGAC ACGCGGCGT GCACACCATG GAGAAGGTTT GGAAAGATA
1381 GTAGGTAGCT TAAACTTCCGG GAG
NCOR3

1 GCAGTCGAGC GGTAACAGGG GAAGCTTGCT GCAGGTAGCT TCCTGCTGA CGAGCGGCGG ACGGGTGAGT
61 AATGTATGGG GATCTGCCCG ATAGAGGGGG ATAACTACTG GAAACGGTAG CTAATACCGC
121 ATAATCTCTC AGGAGCAAAG CAGGGGAACT TCGGTCCTT G CGCTATCGGA TGAACCCATA
181 TGGGATTAGC TAGTAGGTGA GGTAATGGCT CACCTAGGCC ACGATCCCGT GCTGGTCTGA
241 GAGGATGATC AGGCCACTG GAAGCTGACG TCCCTACGGG AGGCAGCAGT
301 GGGGAATATT GCACAATGGG CGCAAGCCTG ATGCAGCCA T GCCGCGTGTA TGAAGAAGGC
361 CCTAGGGTTG TAAAGTACTT TAGTGGAGGA GGAAGGGTAT GATGCTAATA TCATCAACGA
421 TTGACGTTAC CGACAGAAGA AGCAGCGCTC AACTCCGTCG CAGCAAGCCG GGTGATACCGG
481 AGGGTGCAAG CGTTAATCGG AATTACTGGG GTAAAGCGC ACAGACGGCGG TGTAGTAAGT
541 TAGATGTGAA ATCCCGGACC TAAACCTGGG AATTGACCGT CTAGCAGGTC AGCTAGAGTC
601 TTGTAAGAGG GGAGTAGAATT TCAAATCTTT TAGTGGAGA TGGTGAATA GGGAGAATA
661 CCGTGAGCCA AGGGAGCCCA CTGGACAAAG ACTGACGGTC AGGTGCGGGA AGGTTGGAG
721 CAAACAGGAT TAGATACCCT GTGAGTACCC GCTGTAACAG ATGTCGATT TGAAGGTGTT
781 CCCCTAGGAA GTGGCTTTTC GAGCTAACGC GTAAATCGA CGCCTGGGAG AGTACGCGGC
841 CAAGGTAAA ACTCAATGTA ATGGACGGGG GCCCGACAAA GCCGTGGAGAC ATGGCTGTTA
901 ATTCGATGCA ACGCGAAGAA CTTACCTAC TCTGACATC CAGAGAATT ACGAGAGATG
961 CTTGTCGAC CTTGGAAGGT GCTGACAGGG GTGCTGACAG ATGGCGTCTG AGGTTGGAGT
1021 GAAATGTGTTG GTTAAGTCCC GAACAGGACG CAACCCTTAT CTTTGTTGTC CAGCGATTGC
1081 GTCCGGGACT CAAACGACGTC TGCGGTGTAT AAGCGGAGG AAGGTGGGGA TGACGTCAG
1141 TCTATCATGCG CACTACGAGT AAGGCTACAC ACATGTGCTA ATGCCGTATA AAAAGAGAAG
1201 CGACCTCGCG AGAGCAAGCG GAAGTCTCATA ATGACGCTG ATGCGGAGAT TTGAGTCTGA
1261 ACTCGACTCC ATGAGTGCGG AATCGCTAGT AATCGTAGGATA CAGAACTGTA CGGTAATAC
1321 GTCCCGGCGG CTGATACACA CCGCCCGCTCA CACCATGAGG GTGGGTGGCA AAAGAATAG
1381 GTAGCATTAC CTTGGGGAAG GCCGCTA
THKB1

1  GCGTGGATCTA CCTGTCACAG GCACCGAGG GAGCTTGCTC CCTTACGTCA GCGGCGGACG GGTGAGTAAC
61  ACGTGGGAAT CAACTGCATG GAACTGGGTG CTAGACTGCTA CCACGCCGAC GATGGGCTGAC
121  GCTGGAACCG GCAGGTTTGA GCCTGACGTG GACTGGGTCG GGGATCGATG
181  CGACCCCGAG GCACGCTGAGA GGGTGATCGG CCACACTGGG ACTGAGACAC GGGCCGAGCT
241  GAGTCACGTA GGAGGAAAGG GAAGTGGAAT CTGACGCTGAC ATGGAAGGAT
301  GCAGTGCTCGT GCGAGTCTCT TCTACCTGCTG GAGGAGGAGG AGTGGAATTC
361  CAGTGGCGAG GACCTGCGAT CAGGGGTACG TCTACGCTGCT ACGACTGACG
421  TCTTCGCACT GTTGACGGAT CTATTGGAGG GGAGAGGAGG AGTGGAATTC
481  GTGTCACGTA GGAAGAATCG TCGAGACGAC TGGTACGACG GATGAGGGAG
541  TTAGGGGCAG GGTGGGGAAG GGGGATTTCG CTCTACGCTGAC AGTCATGCGA
601  GGAGGAAAGG GAAGAGGAGG AGTGGAATTC CAGTGGCGAG GACCTGCGAT
661  CAGTGGCGAG GACCTGCGAT CAGGGGTACG TCTACGCTGCT ACGACTGACG
721  GAGTCACGTA GGAGGAAAGG GAAGTGGAAT CTGACGCTGAC ATGGAAGGAT
781  CAGTGGCGAG GACCTGCGAT CAGGGGTACG TCTACGCTGCT ACGACTGACG
841  TCTTTCAGCT TCTGCTGACT CTATTGGAGG GGAGAGGAGG AGTGGAATTC
901  CTAGGGGACT TCTGCTGACT CTATTGGAGG GGAGAGGAGG AGTGGAATTC
961  CAGTGGCGAG GACCTGCGAT CAGGGGTACG TCTACGCTGCT ACGACTGACG
1021  CAGTGGCGAG GACCTGCGAT CAGGGGTACG TCTACGCTGCT ACGACTGACG
1081  CAGTGGCGAG GACCTGCGAT CAGGGGTACG TCTACGCTGCT ACGACTGACG
1141  CAGTGGCGAG GACCTGCGAT CAGGGGTACG TCTACGCTGCT ACGACTGACG
1201  CAGTGGCGAG GACCTGCGAT CAGGGGTACG TCTACGCTGCT ACGACTGACG
1261  CAGTGGCGAG GACCTGCGAT CAGGGGTACG TCTACGCTGCT ACGACTGACG
1321  CAGTGGCGAG GACCTGCGAT CAGGGGTACG TCTACGCTGCT ACGACTGACG
1381  CAGTGGCGAG GACCTGCGAT CAGGGGTACG TCTACGCTGCT ACGACTGACG

46
THKB2

1  TGCAGTCCAG CGGACAGATG GGAGGTTGCT CCCTGATGTT AGCGGCGGAC GGGTGAGTAA
61  CACGTGGGTTA ACCTGCTGTG TAAGACTGGGA TAACTCCGGG AAACCAGGCTG TAATAACCGGA
121  TGGTTTGGTG AACCAGATGG TCCAAACATA AAAGGTGAGCCC TCGGCTACCA GTTACAGATG
181  GACCTGAGA GGTTGTACCG CCACACTGGG ACTGAGACAC GCACCAGACTC CCTACCGGAG
241  GCAGCAGTAG GTAATCTTCG CGAATGGCAGAA AAGTCTGACG GGAGGACGAGA CCGTGAGTGGT
301  ATGAGGTTTT TCGGATCTGA AGCTCCTTTT GTTAGGGAAGA AAAAGTACCC GGTCGGAATAG
361  GGCGGTACCT TGACGCGTACC TACACAGAAA GCCACGACAT ACTACGTGCC AGCAGCGCGG
421  GAAATACGTTA GGTGGCAAAGC GTTTGCCGGA ATTTATGCGTTT GAAAGGCTC CGCCAGCGG
481  TTCTTAAGTC TGATGTGAAA GCCCCGGCTCA ACCCGGGGCA GGTACATTGGG AAGACTGGGA
541  ATTGAGTAGA AGGAGGGGAAGT GGAATCGCAGA TCGCACTCGG GCAGCAGAAGC
601  GGAGGAACAC GAGTGCGAAC GGGACTCTGC TGGTCTGTA A CTCGAGTGCAGA GAAGGATCTG
661  CGTGGGGAGC GAACAGGATT AGATACCCTG GTAGTCCAGC CCGTAACACG TGAGTGCTAA
721  CTTCCGGGCTC CCGTAATACG ACTACGTGCC AGCAGCCGCGG
781  GTGTAGGCGG GCTTAGTGGTG CAGCTAAAGC ATTAAGCAGA CCGCCTGGGG
841  AGTTACGCTCG CAAGACTGAA ACTCAAAGGT TAATCGGAGG TGGCACCGGG GCACGGCACA
901  AGCGGTTGAG CATGTGGTTT AATTCGAGGC ACGCGAAGAC CCGTAACGAC GTCTGTGACAT
961  CCTCTGACAA TCCTAGAGAT AGGACGTCG CTTCCGGGGC TGGTACAGA CCGTGACATG
1021  GTGTGGTCGT TCGCTGGTGTG TGGATGGTTT GGTTAAGTCC CGCAACAGAG GCAACCTGTG
1081  ATTTTTCCATTCA GTGGGACACT CTAAGGTCAC CGCGGATGAC AAAACGGGAG
1141  AAGGTGGGGA TGACITCAA TACTCATGCC CCTATGACC TGGGCTCACG ACCTGCTACA
1201  ATGGACAGAA CAAAGGGCAG CGAACCAGCG AGGTTAGGAC AATCCCAATA ATCGTCTCTC
1261  AGTTCCGATC GCAGTGCTAC SGCGACTGC GTGGACTGGA AATCGTCAGT GAAACCGGAT
1321  CAGCATGCCG CGGTGATATA GTGGGCTGGG CTGTGACACA CCGCCTGGCA CACCAGCAGA
1381  GTTTGTAACAC CCCGAAAGTCG GGAGGTAAC
NCKP1

1 TGCAGTCGAG CGGTAGCACA GGGGAGCTTG CTCCTGCCGA GACGAGCGGC GGAAGGGAGTGA
61 GTAATGTCTG GAAACTCTCC TGGAGGGGGAAGAATGAA CTTGGGCCCT CTTGCCATCA GATGTGCCCA
121 GATGGGATTAG TCTAGATTGG CTGGCTAGGCC CTACTCTAGG CCAGGATACCA GACTCCTACG GGAGGCAGCA
181 GTGGGGAATA TTGCACAATG GCGCAGAAGCC TGATGCAGCC ATGCCGCGTG TGTGAAGAAG
241 GCTTCGCGGT TGTAAGGCAC TTTACGCGGAAGAAGGTGGT GTGARCTTAA TACGTTAAC
301 TACGGTGCCGA AGAAGCATG GCCAGCAGAC CCCTGGACGA AGACTGACG TCACTGTACC GACCTGACG
361 GTCTTAGAGA GGGGGTAGAA TTCCAGGTGT AGCGGTGAA ATCTGTGGAA TGGAGGGAG
421 AGAAACTATTGA ATTACATACC TCTCTAGTCC AACGTCTTAA CAATGTCCGA TTGGAGGTGTG
481 TGGCCCGTGA GTGGGAAAGG CGGTAAACGA GGCCGGATTG GCAACGTGCA TCTGGAGGAA
541 GCAAGGGTCGA AGCCTTAATC GGAATTACTG GCCGTAAGCC GGGGATGGCTG TGTGAAGAAG
601 TCTTGAGAGA GGGGGTAAATG GCAAGCTAGAG TCTAGTCTG AATCTGGGGG GGGGAGCTTG
661 GTAATGGGCTT TCGCTGGAGA GCTCAACCTG GGAAGGCTAGG CAATGTCCGA TGTGAAGAAG
721 AGAAACTATTGA ATTACATACC TCTCTAGTCC AACGTCTTAA CAATGTCCGA TTGGAGGTGTG
781 TGGCCCGTGA GTGGGAAAGG CGGTAAACGA GGCCGGATTG GCAACGTGCA TCTGGAGGAA
841 GCAAGGGTCGA AGCCTTAATC GGAATTACTG GCCGTAAGCC GGGGATGGCTG TGTGAAGAAG
901 TAATGATGCT CAAAGCTTATG AGCTTTAGAGA GGGGGTAAATG GCAAGCTAGAG TCTAGTCTG
961 GATTTTGAGA CTCCCGTAAAG TCTGTACGCA ATCCGGAGA ACCTGCCAGG GGAAGGCTAGG
1021 TAAATCTGTTT GGTGGTACG GCGGATACCA GCAAGCTTATG AGCTTTAGAGA GGGGGTAAATG
1081 GCAAGGGTCGA AGCCTTAATC GGAATTACTG GCCGTAAGCC GGGGATGGCTG TGTGAAGAAG
1141 GTCAATCTTT TCTCTAGTCC AACGTCTTAA CAATGTCCGA TTGGAGGTGTG GCAACGTGCA TCTGGAGGAA
1201 GCGCCTCGAG CAGAGGAAGG CGACCTTATA AATGGCTTAT CACCCGTAAGGTTGTTG GCAACGTGCA TCTGGAGGAA
1261 AACCTGACTC CATGAAATCG GGATCGCTGAGAAGGCTAGG CAATGTCCGA TTGGAGGTGTG GCAACGTGCA TCTGGAGGAA
1321 GTGGGGAATA TTGCACAATG GCGCAGAAGCC TGATGCAGCC ATGCCGCGTG TGTGAAGAAG
1381 GGTACGCTTAA CCTCGGGAG GCGGCTAC
NCKP2

1 TGCAGTCGAG CGGACAGATG GGAGCTTCTG CTCTGATGTT AGCGG CGGAC GGGTGAGTAA
   61 CACGTGGGTA ACCTGCTGTG AGAAGTGGGA TAATCCGGGG AACCCGGGGA TAATACCGGA
   121 TGGTTGTGGG AACCGCATGG TTCAAACATA AAAGGCTTGGC TCGGCTTACCA TCTACAGATG
   181 GACCCGGCCGC GCATTAGCTA GTTGGTGAAGG TAAAGGGCTC CCCAGGCA TCAGCGTACGC
   241 CGACCTGAGA GGGTGATCGG CCAGACTGGG AAGACTGGGAC GTTGAAGGAGG AGGTGGGGAT
   301 GCAGCAGTAG GGAATCTTCC GCAATGGACG AAAGTCTGAA GGGAGCAACGC CGCGTGAGTG
   361 ATGAAAGTTT TCGGATCGTA AGGTCTGTTT GTTAAGGAAAG AACAGTACCA GTTCGAATAG
   421 GGCGGTACCT TGGACGGTACC TAACCCAGAAA CCCAGGCTCA ACTAGTGCCC AGCGGCGCGG
   481 GTAATACGTA GGTAAGGAGG AGTGGAGATG CACAGTGGC AATTATTGGGC TTAAAGGCTC CTCAGGCGGT
   541 TTCTTAAGTCT TGGATGCGAA GGGCTCCGCC CTAAACCGGA GGGTATTGAA GAACTGGGGAA
   601 ACTTGAATGG GAAAGAGGAGG AAGGGCTCAGG TACAGTGCAGG GTGAAATGGC GTAGAGATGT
   661 GGAGGAACAC CAGTGCCGAA GGGCACTTCA TGGTCTGTA GCAGCGCTGA GGAGCGAAAG
   721 CGTGGGAGCG GAAACGGATT AGATACCCC TGGTACCGCG CCGTAAAACAG TGAGTGCCTA
   781 GTGTTAGGGG GTTTTCGCCA CTAGTGTCAG CAGCTAAAAC ATTAAGCAGT CCGGCTGGGG
   841 AGTGAGGATG CAAAGCTGAA ACTCAGGAGG GTGGGACGGGG GCAGCAGCAA GCGGTGGAGG
   901 ATGGGTGTTA ATTCGAAAGC ACGGCAGAGG CTTTACCAGG CCTTGACATG CTCTGACAAT
   961 CCTAGAGATA GGACGTCCC TCCGCGGGCA GAGTGACAGC TGGTGCATGG TTTGGTCGCA
  1021 CTCGTGTCTG GAGATGTGGG GTTAAAGGGA GCAACGAAGGG CAACCCTTGA CTCTGTTGCA
  1081 CAGGATTTAC TGGGGCACTC TAAGGGTACT GCGGGTACA ACGCGGAGAG AGGTGGGAT
  1141 GACGTCAAAT CATCATGGCC CTTATAGGGT ACGCTACACA CGTGCTACA TGGACAGAAC
  1201 AAAGGGCAGC GAACCGGGCA GGGTAAAGCGA ATCCCACAAA CTGGTCTTACA ATGGGAT
  1261 CAGTCTGCAA CTCGACTGCG TGAAGCTGGA ATCGCTAGTA ATCGCGGATC AGCATGCCGC
  1321 GCGGTAAGAG TTCCGGGGCC TGTGATAGG ATGGAGGAGG ACCACGGGAG GTCGAGAAC
  1381 CGGAAGTCCG TGAGGTAACC TTTAGGGACCC A
TLPR1

1 GCAGTCGAGC GGTAACAGGA GAAAGCTTGC TTTCTTGCTG ACGAG CGGCG GACGGGTGAG
   61 TAATGTATGG GGATCTGCCS GATAGAGGGG GATAACTACT GGAACCTGGT GCTAAATACG
   121 CATAATGTCT AGCGAGCACA AACAGGCGGG TTTCGACCTT GCACTATCGG ATGAACCCAT
   181 ATGGGATAGG TAGTAGGGT GGTAAAGGCC TCACCTAGGC GACGATCTCT AGCTGGTCTG
   241 AGAGGATGAT CAGGCCACCT GGGAATGAGA CACGCGCCTG ACTCTACAGG GAGGACGAG
   301 TGGAATATGG TGCAAAATGG GCGCAAGCCT GATGCAGGCG TGCTGTTTT GAGGAAGGAG
   361 CATTAGGTTA GTAAAGTACT TTCACGCTG AAAGAGGTGA TAAAGTATAC ACCCTCTATCA
   421 ATGGACGTTA CCCGCAGAGG AACAGAGCCGC TAACTCCGGG CGAGCAGCCG CGTAATACG
   481 GAGGGTTCGA GATGTAATCG GAATCTCTGG GCGTAAAGCC CACCGAGGCG GTCAATCCAG
   541 TCAGATGTGA AAGCGCCAGG CTAACTCTGG GAAATTCGAG TGAGGCTTTT TGGCTAGAGT
   601 CTGTAGAGGG GGAACTTGGAG GAATCTCTGG GCGATGCTGA GTGGACTTTT GCAGGTTTTT
   661 ACCGCTGGGC AAGCCACGGCC CCTGGGACCA GACTGAGCCT CAGGTGGGAA ACGGCTGGGA
   721 GCAAAACAGGA TTGATACCC TGGTGAACCA CGCTGAAAAC GATGCTGATT TAGAGGTTGT
   781 GGTCTTACCG CGTGGCTTTT CAGACCTGGC ACGCTCGAGG GATGACCGCC ACCCAGGCGG
   841 GCAAGCTTAA AACTCAAATG AATTGAGGGG GCGCCAGACA AGCGTGTTGG AGTGAGGTGT
   901 GATTCAGGCT AAGCGCAAGG GGGCAGAGAA TATCTCCTGG ACAGGCTGTA ATGGGCTTTT
   961 AGAGGAGTGC CTTCGGGAAA CTCTGACAGG GTGCTGCATG GCTGTCGATG GCTGCTGTTT
  1021 TGAATATGGT GGTGTAATCC GGAACGCAGC GCAACCCCTA TCTTTTGTTG CCAGCAGGTA
  1081 ATGGTGGGAA CTCAAAGGAG ACTGCCGCTT GATAACCCGA GGAAAGTGGG GATGACGTCA
  1141 AGTCATCATG GCCCTTACGA GTAGGCTTAC ACAGTGCTA CAATGGCAAG TACAAGAGA
  1201 AGCGACCTCG CGAGACAGCG GGAACCTCAT AAAGCTGTGCA GTAGTCCCGG ATGGAGTCTG
  1261 CATCATCTG CTAGTACAGT GGAATCGGTA GTAATCTGTA AAGCTGTGCA TACCGAGTAT
  1321 AGTTCCCCGG GCCTTGTACA CACCGCCGGT CACCCATGG GATGGGTTTG CAAAGAAGT
  1381 AGGTAGCTTA ACCTTCCGGG
TLPR2

1 GCAGTCGAAC GTAAACAGGA AGCAGCTTGC TGCTTGCTGT ACGAGTGCCG GACGGGTGAG

61 TAATGTCTGG GAAACTGCCT GATGGAGGGG GATAACTACT GAAACCGGGA TCTAACACCG

121 CATAAACGTCC CAAGACCAAA GAGGAGGGCC TTTGGGCTAC TGCCCATCGG ATGGGTCTTG

181 AGAACCCACT GAGCTACTGAG CAGGATCCAG ACTCCTACG GAAGGCCGCA

241 AGAATGCTTA CCCGCAGAAG AAGCACCAGG CAACTTCCGT CCAGCAGCCG CTGAAATACG

301 GAGGGTGCAA GCCTTTAGAT GCCTAAAGAG CACCGAGGCC GACTCCTACG GAGGGTGAGA

361 CAGCATTTG CTAAGATCTGT CTTAGCTCGG AGGAAAGGGG GAAGGTTAAT AACCTTGTCG

421 ATTGACGTTA CCCGCAGAAG AAGCACCAGG CAACTTCCGT CCAGCAGCCG CTGAAATACG

481 GAGGGTGCAA GCCTTTAGAT GCCTAAAGAG CACCGAGGCC GACTCCTACG GAGGGTGAGA

541 TGGGATGTGA ATCCCCGGG CTCAACCTGG GAACTGCAT TCGAAACTGGC AGGCTAGAGT

601 CTTGTAGAGG GGGGTAGAAT TCCAGGTGTA GCGGTGAAAAT GCGTAGAGAT CTGGAGGAAT

661 ACCGGTGGCG AAGCGGCGCC CCTGGAAACCA ACCTACGCT CAGGTGGCGA AGCGCGGGGA

721 GCAAAACGGA TTAGTACCC TGTAGTCCCA CGCCGATAAC GATGTGACT TGGAGGTGTG

781 GCCCTTGGG GCTGGGTCTTCC GAGCTAACGC CGATTAGTCG ACCGCCCTGG GAGTACGGCC

841 GCAAGGTTAA AACTCAAAATG AATTGACGGG GGCAGGCGCA AGGCGGTGGG GATGTGGTTT

901 AATTCCGATGC AACGGGAAAG AACTCTACTT GCTGGACAT CGAGAAGACT CAGGAGAGAT

961 GCTTGTTGTC CTCCGGGAAT GTGGGTACAG GTGCTGTGCA GCTCGTGTGT GCTCGTGTGT

1021 TAGAAATGTG TGGTTAAGCTT CCAAGCCGAC GCAACCTTAA CTCCTTGTGT CCAGCAGGTC

1081 GCAGGCGGGA TCAAGGAGAG CAGCCGATTG TAAAATGGGA GAGGTTGGG GATGACGTCAA

1141 CTCTGATCCG ACCGACTCAG TGGGGCTACA CACGACTGCA AATGGGCGCAT ACACGAGAAA

1201 GCCGACCTGCG GAGACAGACC GAGGAGTACAT CAAAGCGCGT GGTGGGTGAC GTGCTGTGCT

1261 AACTCGACTC AGAATGCTTG GTAAGGGCTG TCAAGATGCA ACGGGAAGAT

1321 CGTCCAGGGG CTTGTACAC CCAGGCAACT AAGCGGCGCA AACGCGGAGG AGTGGGTTTC AAAAGAAGTA

1381 GCTTGCTTAA CTTCCGGAG GGGCGCT
TLPR3

1   TGCAGTCGAC GGTAACAGGA AGCAGCTTGC TGCTTCGCTG ACGAG TGGCG GACGGGTGAG
   61  TAATGTCTGG GAAACTGCCT GATGGAGGGG GATAACTACT GGAAAACGGTA GCTAATACCG
 121  CATAACGTGC CAAGACCAAA GAGGGGGACC TTCGGGCCT C TTGCCATCGG ATGTGCCCAG
 181  ATGGGATTAG CTAGTAGGTG GGGTAACGGC TCACCTAGG C GACGATCCCT AGCTGGTCTG
 241  AGAGGATGAC CAGGCCCACGT GAACTGAGA CACGGTCCA G ACTCCTACGG GAGGCAGCAG
 301  TGGGGAATAT TGCACAATGG GCGCAAGCCT GATGCAGCC A TGCCGCGTGT ATGAAGAAGG
 361  CCTTCGGGTT GTAAAGTACT TTCAGCGGGG AGGAAGGCGA TGAGTTAATT AACCTTGTCG
 421  ATTGACGTTA CCCGCAGAAG AAGCACCAGC GCACCTCCGT CCAGCAGCCG CGGTAATACG
 481  GAGGGTGCAA GCGTTAATCG GAATTACTGG GCGTAAAGCG CACGCAGGC CGTCTGTCAG
 541  TCGGATGTGA AATCCCCGGG CTCAACCTGG GAACTGAGA T CCGCTTACAT CACGGTGTTT
 601  CTTTGGTGC CTCCGGGAAC TCTGAGACAG GTGCTGCAT G GCTGTCGTCA GCTCGTGTTG
 661  ATGGTGTCTG CTAAGAACAGG TCTTCAATAG GGCAACGTCA TGGAGTTTGA AATGGGTTTC
 721  TCAAGAAGTTA AATCTCAATAG AATGGAGGAG GGCCCGCACA AGCGGTGGAG CATGTTGTTT
 781  CTACCCCAATCC ACCCAAACTCC ACCCGCAGCT ACCTTACACT GGAAAGGAGG GAAGGTGGGG
 841  GACGGTGAA TGGGTGATGC AATCTCAATAG GGCAACGTCA TGGAGTTTGA AATGGGTTTC
 901  TCAAGAAGTTA AATCTCAATAG AATGGAGGAG GGCCCGCACA AGCGGTGGAG CATGTTGTTT
 961  GCTGCTTACGT CCTCAGCAAC GACTTGACAG TGCTGAGGAG GTGCTGCTGCA GTGCTGCTT
1021  TCAAGAAGTTA AATCTCAATAG GGCAACGTCA TGGAGTTTGA AATGGGTTTC
1081  ATGGTGTCTG CTAAGAACAGG TCTTCAATAG GGCAACGTCA TGGAGTTTGA AATGGGTTTC
1141  GTGCTGTCGCT CACTCGTACG TGGGCTGAGG GAAAGGAGG GAAGGTGGGG
1201  GACGGTGAA TGGGTGATGC AATCTCAATAG GGCAACGTCA TGGAGTTTGA AATGGGTTTC
1261  AACTCGACTC CATGAAGTCG GAATCGCTAG TAATCGTGGG TACACTGACGT ACTAAGTACC
1321  CGCTGACTG GCTTACAC ACGCCCGTTC ACGCATGAG GTGGGTTGGG AAAGAAGAAGT
1381  AGGTA Cait ACGGCTCGGG AGGGCGCT