CLASSIFICATION OF A SPECIES OF *ERWINIA* FROM THE OCONALUFTEE RIVER, GREAT SMOKY MOUNTAINS NATIONAL PARK

A thesis presented to the faculty of the Graduate School of Western Carolina University in partial fulfillment of the requirements for the degree of Masters of Science in Biology

By

Robert Pollock McKinnon

Director: Dr. Seán O'Connell

Associate Professor of Biology

Department of Biology

Committee Members: Dr. Katherine Mathews, Department of Biology
Dr. Maria Gainey, Department of Biology

June 2017

ACKNOWLEDGMENTS

I would like to thank my committee members and director for their encouragement. In particular, I would like to thank Seán O'Connell for his patience, guidance, and his ability to create time where none exists. Thank you to Kathy Mathews for assistance and patience with phylogenetic analyses. Thank you to Maria Gainey for introductory lessons in bioinformatics and whole genome analysis. Thank you to Darby Harris for editing and notes. Thanks to Jamie Wallen and Dr. Grant for TEM imaging. Thank you to Kacie Frasier for teaching me how to make various media.

I would also like to thank my fellow current and former graduate students, particularly but in no specific order Cameron and Jessica Duke, Kyle Corcoran, M.J. Michaels, Jessica McLamb, Sam McCoy, Thom Green, Kwame Brown, Tori Carlson, Sarah Britton, and Hannah Meeler. I' would also like to thank any teacher I assisted under, any student I taught, and anyone else who I may have learned from either intentionally or not. Thank you to Tom Scharpling, Dan Harmon, Chris Gethard, Connor Ratliff, J.D. Amato, Kevin T. Porter, Demi Adejuyigbe, Nick Wiger, Mike ""Mitch" Mitchell, Scott Auckerman, Alie Ward, Georgia Hardstark, and Lauren Lapkus and all of the therapeutic content you've produced.

Finally, thank you to my friends and family who've taught, supported, and shaped me in countless ways.

TABLE OF CONTENTS

List of Tables

List of Figures

List of Abbreviations

ATBI: All Taxa Biodiversity Inventory
BLAST: Basic Local Alignment Search Tool
CI: Consistency Index
IJSEM: International Journal of Systematic and Evolutionary Microbiology
GSMNP: Great Smoky Mountains National Park
GTR: General Time Reversible
KP: Kephart Prong
ML: Maximum Likelihood
MSA: Multiple Sequence Alignment
NP: Natural Products
OR: Oconaluftee River
PCR: Polymerase Chain Reaction
RDP: Ribosomal Database Project
RI: Retention Index
TBR: Tree-bisection-reconnection

Abstract

CLASSIFICATION OF A SPECIES OF *ERWINIA* FROM THE OCONALUFTEE RIVER, GREAT SMOKY MOUNTAINS NATIONAL PARK

Robert Pollock McKinnon

Western Carolina University (May 2017)

Director: Dr. Seán O'Connell

As part of the All Taxa Biodiversity Inventory (ATBI) project from Great Smoky Mountains National Park (GSMNP), stream water was collected from near the Oconaluftee Visitor Center, and bacteria were cultured from it. One of the 30 bacterial isolates was then selected and subjected to a taxonomic investigation using a polyphasic approach of growth-based and DNA-based methods. The isolate was found to be a facultative anaerobe, Gram-negative, rod-shaped in cell morphology, and motile. The species showed a temperature growth range from 4℃-30℃, pH range of 4-7, and salinity tolerance of up to 3%. Metabolic analysis of the isolate indicated the strain to be catalase positive, oxidase negative, and it had the ability to ferment glucose and arabinose. Analyses using the Ribosomal Database Project (RDP) showed the isolate to have the highest 16S rRNA gene sequence similarity with *Erwinia aphidicola* at 92.7%. Genome sequencing and annotation indicated twelve conserved signature indels (CSIs) present, with 65 drug resistant related genes, eight transposable elements, fifty genes with a phage origin, and one pathogenesis related gene. Maximum likelihood and parsimony models using the 16S rRNA gene were used for phylogenetic analysis of the isolate and its closely related species in the genera *Pantoea* and *Erwinia*. Results indicated that the isolate is a strain of *Erwinia billingiae*,

which was corroborated via sequencing of the full genome. *Erwinia billingiae* is an epiphyte associated with *Rosaceous* plants, and has been examined as a potential biological control to *Erwinia amylovora* the species that causes fire blight.

Chapter 1: Introduction

Although most are invisible to the naked eye, it's estimated some 10^{30} microbial cells are present on the Earth (Whitman et al. 1998). Bacterial species diversity is also high with an estimated tens of millions of microbial species in ten grams of soil (Whitman et al. 1998), some two million species within the global oceans (Curtis et al. 2002), and latest predictions are shown to estimate over two thousand different phyla within the bacterial domain (Yarza et al. 2014). Between the sheer numbers of microbial species and their quickly evolving nature, the challenges present in fully categorizing and classifying microbial species and communities have been well documented (Quince et al. 2008, Sloan et al. 2007, Cohan 2001, Koeppel et al. 2013). This incomplete knowledge of species characteristics and diversity inhibits microbial ecologists from using ecological concepts and tenets to further advance the fields of industry, human health, bioremediation, and biotechnology, among others (McFall-Ngai 2015).

The current era of microbial systematics began in 1977, when Carl Woese and colleagues at the University of Illinois began to apply the methods of molecular systematics to 'prokaryotes' (Woese et al. 1977). Woese used the small-subunit rRNA gene (16S rRNA of bacteria and 18S rRNA of eukaryotes) as a homologous marker for phylogenetic reconstructions (Fox et al. 1977). The use of the 16S rRNA gene in conjunction with the exponential growth of sequencing technology has allowed for many different studies in the fields of microbial biodiversity and taxonomy without culturing, e.g., Smith et al. (2015), Mahnert et al. (2015). One drawback to these techniques is that the phenotypic traits of the sequenced organisms and their metabolic processes can only be inferred and often go unknown. Prior to the advancements in sequencing

technology, taxonomic research was done through the cultivation and isolation of natural samples with growth media. Isolated cultures were then classified/diagnosed through their size, morphology, physiology, and biochemical characteristics, using such resources as Bergey's Manual of Determinative Bacteriology as a guide (http://www.bergeys.org/).

The International Journal of Systematic and Evolutionary Microbiology (IJSEM) currently defines a prokaryotic species as "a category that circumscribes a (preferably) genomically coherent group of individual isolates/strains sharing a high degree of similarity in (many) independent features, comparatively tested under highly standardized conditions" (Stackebrandt et al. 2002). Or more specifically, a prokaryotic species is considered to be a group of strains (including the type strain) that share a certain degree of phenotypic consistency, show 70% of DNA–DNA binding, and have over 97% similarity in 16S rRNA gene-sequence identity (Vandamme et al. 1996).

As sequence similarity of a complete 16S rRNA gene diverges further from known strains, the chance of uniqueness for an isolate increases. The general "tier system" for sequence similarity laid out by IJSEM shows this negative correlation between sequence similarity and uniqueness of a strain. An isolate with greater than 97% similarity to another is part of the same species, an isolate at 95% or greater similarity is a species within the same genus, a sequence similarity of 90% or greater represent species within the same family, and sequence similarity greater than 80% indicates a species at the same order level (Stackebrandt & Goebel, 1994). The complete 16S rRNA gene indicates where an isolate falls taxonomically, which further informs direction of testing in regards to 70% DNA-DNA hybridization and phenotypic characterizations. The 70% DNA-DNA hybridization test indirectly measures the overall

similarity between two genome sequences (McCarthy & Bolton, 1961). Next generation sequencing technology and the advancing ease of whole genome sequencing has provided a substitute for the more labor-intensive and less informative DNA–DNA hybridization for studies in taxonomy of bacteria and archaea (Kim et al. 2014).

Whole genome sequencing has allowed for studies in taxonomy and classification to be conducted at a deeper level, further parsing out relationships at just not a single gene, but at the genomic, proteomic, and transcriptomic levels. Genome sequence data has enabled the detection of conserved molecular characteristics shared by evolutionarily related groups of organisms. One particular group of conserved molecular characteristics, which have been utilized to great effect in prokaryotic taxonomy are conserved signature insertions/deletions (CSIs) present in widely distributed proteins (Gupta, 2014). CSIs are insertions or deletions that are uniquely present in a related group of organisms. The most parsimonious explanation of the presence of the CSI in a related group of organisms is the existence of a common ancestor in which the genetic change leading to the CSI occurred, and which was subsequently inherited by all of its various decedents leading to synapomorphic characteristics (Gupta, 2014). These synapomorphic characteristics can further delineate taxonomic rank for example, the 69 CSIs that have been found in the "Enterobacteriales" family, shows support that the 60 genera comprising the family should be broken into seven distinct families (Adeolu et al. 2016). Just as in eukaryotic species analysis and classification, each taxonomic rank further clarifies the phenotypic traits that should be tested for a more complete classification.

Such resources as Bergey's Manual of Determinative Bacteriology show what phenotypic traits are consistent among taxa of a certain genus, with some consistency also occurring at the

family or order level (Van Belkum 2006). For example, if an isolate's 16S rRNA gene sequence shows that it is most closely related to species in the genera of *Pseudomonas* and *Xanthomonas*, testing for production of the unique yellow pigments known as xanthomonadians, which are novel to *Xanthomonas* as well as testing for more specific growth factors, will further delineate where an isolate falls between those two genera (Van Belkum 2006). Taxonomic ranking becomes more specific as one tests more features, as do the suite of tests available to clarify between taxa, for example, detection of pigment production, testosterone degradation, reduction of nitrate and denitrification, acid production from carbohydrates, and ring fission mechanisms for cell division, which are just a few of the special phenotypic characteristics within the *Pseudomonas* genus (Van Belkum 2006).

The current methodology for defining a prokaryotic species involves obtaining the 16S rDNA sequence (and other genes as necessary), gathering phenotypic data (e.g., organics used, enzymes produced, fatty acids used in the cell membrane), a deposit of cultures into two culture collections (e.g., American Type Culture Collection, Deutsche Sammlung von Mikroorganismen und Zellkulturen), and the publication of a species account into IJSEM (Stackebrandt, et al. 2002). This polyphasic approach towards classification and naming of bacteria and archaea is causing an ever expanding gulf in regards to available sequence data and the characterization of phenotypic traits for the organisms that these sequences have been derived from. In fact, the rapidly expanding publicly available sequence databases of the 16S rRNA gene have the number of entries currently exceeding 4 million, which in conjunction with the slow going approach of culture work, combine to have about twelve thousand or 0.3% of available species having been

classified and named (Yarza et al. 2014). These figures demonstrate a great need for cultivation studies to expand the number of well-classified bacteria available in biology.

An ecological and taxonomic project local to western North Carolina and eastern Tennessee is that of the All Taxa Biodiversity Inventory (ATBI) taking place in Great Smoky Mountains National Park (GSMNP). The ATBI is a project of Discover Life in America and is seeking to inventory the estimated 100,000 species of living organisms in GSMNP. As various studies have indicated, 100,000 is likely an underestimate for microbial species alone (Whitman et al. 1998). Assessing microbial diversity within areas of the GSMNP can act as ecological indicators for water and soil quality (Garrido et al. 2014, Ndaw et al. 2009) and provide a unique window to the most diverse domain in one of the more biologically diverse regions on Earth. The purpose of this study was to isolate bacterial species from GSMNP, select one for further taxonomic study, and classify a presumptive new species of bacteria.

Chapter 2: Methods

Section 2.1: Isolation and Preliminary Classification of Bacterial Species

Section 2.1.1 Sampling Technique and Location

Samples were taken from two sites in GSMNP, March 26, 2015 and August 20, 2015. Three samples were taken at a site near the Oconaluftee Visitor Center (UTM 17S 0291211, 3931770) in March and one was taken at the Kephart Prong trail (UTM 17S 0286057, 3940540) the following August. Samples were aseptically collected from stream water, by immersing a sterile centrifuge tube in the water and filling to 80% volume. Each sample was collected within different micro-habitats within the stream (calm, rippled, and shore line) and each upstream from the previously collected sample to minimize sediment collection. All samples were placed into a cooler on ice packs and returned to the lab where they were kept at 4□ until culture work began (O'Connell et al. 2007).

Section 2.1.2: Spread Plating & Isolation of Culture

Cultures were obtained by serially diluting environmental samples in sterile water and spread-plating the dilutions onto R2A and TSB media plates (S. O'Connell personal communication). Plates were inverted, incubated in the dark at room temperature and assessed for growth after one week, at which point 20 colonies from the Oconaluftee samples and 12 colonies from the Kephart Prong Trail samples were chosen for isolation. Each colony was streaked for purity, and this process was repeated 4-6 times until pure cultures comprised of one species each had been obtained.

Section 2.1.3: DNA Extraction and PCR

DNA was obtained from each isolate by using a Mo Bio Ultraclean Microbial DNA Isolation kit (Mo Bio, Inc., Solana Beach, CA). Polymerase chain reaction (PCR) conditions to amplify ca 550 bp of the 16S rDNA from each isolate consisted of the following: 50 uL total volume in nuclease-free water with final concentrations of 1x Mastermix chemistry (Promega) (Madison WI), 0.25 uM of bacterial specific 341F and 907R primers (Casamayor et al. 2000), and genomic DNA. PCR was conducted using a "touchdown" approach (Casamayor et al. 2000) on a Mastercycler Personal thermal cycler (Eppendorf) with initial denaturation for 5 min at 94℃, followed by 30 PCR cycles consisting of; 1 min denaturation at 94℃; 1 min annealing at decreasing temperature (beginning at 65℃ for two cycles, dropping 1℃ at each cycle for ten cycles, and ending at 55℃ for 18 cycles); 3 min elongation at 72℃, and a final elongation for 7 min at 72℃ with a sample hold at 4℃. Genomic DNA extracts and PCR products were screened in 1% agarose gels, and PCR products were cleaned using Montage spin filters (Millipore, Inc., Bedford, MA) prior to DNA sequencing.

All 16S rDNA PCR products were then amplified in a sequencing PCR reaction using primer 907R with a BigDye Terminator Version 3.1 Cycle Sequencing Kit (Applied Biosystems, Foster City, CA). The PCR products were cleaned using Autoseq Sephadex G-50 spin columns (Amersham Biosciences, Piscataway, NJ) and analyzed by a 3130/3130xl automated capillary DNA sequencer (Applied Biosystems).

Sequences were compared to sequences of clones and isolates contained in the Ribosomal Database Project (RDP) website (Cole et al. 2014). Sequence quality was assessed with Finch TV (Geospiza, Inc.) and the Classifier program of RDP was used to classify the isolates from the

level of phylum down to genus. The results included confidence values for placing each sequence into a taxonomic rank. The SeqMatch program of RDP was used to find the top five most similar species in the database. Search parameters for SeqMatch used type strains of isolates, good quality sequences, and length of sequences both above and below 1200 nucleotides. Type strains have already been classified as microbial species and have known phenotypic traits and they served as a reference for each species.

One of the twenty isolates was chosen for further study based on its unique sequence, sequence quality, and sufficient growth conditions. The isolate of choice was sent to GeneWiz labs (South Plainfield, New Jersey) in order to obtain the full 16S rRNA gene sequence. This sequence was used in the phylogenetic analysis below and the isolate was sent out for a full genome sequencing study.

Section 2.2: Classification and Characterization of Bacterial Species

Section 2.2.1: Phylogenetic Analysis

Section 2.2.1.1: Taxa Sampled**.**

A preliminary tree was constructed to determine what taxa should be used for phylogenetic reconstruction. Twenty closely related 16S rRNA gene sequences were retrieved from the RDP program SeqMatch (http://rdp.cme.msu.edu/SeqMatch/SeqMatch_intro.jsp) under parameters of type strains, isolates, sequences greater 1200 base pairs, and high quality sequence. The taxa sampled, sequence similarity (as determined by RDP), and GenBank accession numbers are shown in Table 1.

Table 1: List of taxa used for preliminary phylogenetic tree reconstruction. The twenty species with highest 16S rRNA gene sequence similarity to the isolate, based on the RDP SeqMatch program as of October, 2016, listed with percent sequence similarity, and GenBank accession numbers.

Species Name	**Gen Bank Accession**	**Sequence Similarity (%)**
Erwinia aphidicola	FN547376	92.7
Erwinia persicina	Z96086	92.5
Erwinia tasmaniensis	AM055716	91.5
Erwinia rhapontici	AJ233417	91.3
Erwinia toletana	FR870447	90.6
Pantoea dispersa	DQ504305	90.8
Erwinia amylovora	AJ233410	89.4
Pantoea eucrina	EU216736	89.4
Pantoea septica	EU216734	89
Erwinia piriflorinigrans	GQ405202	88.7
Buttiauxella noackiae	AJ23340	88.2
Erwinia mallotivora	AJ233414	88.1
Buttiauxella izardii	AJ233404	88.1
Enterobacter ludwigii	AJ853891	88
Buttiauxella warmboldiae	AJ233406	87.3
Enterobacter cancerogenus	Z96078	87.2
Pantoea agglomerans	AJ233423	87.1
Enterobacter asburiae	AB004744	87.1
Kluyvera intermedia	AF310217	86.9

Sub Section 2.2.1.1 Preliminary Phylogenetic Analysis.

Sequences were aligned with Clustal Omega into Nexus format (http://www.ebi.ac.uk/Tools/msa/clustalo/). Phylogenetic trees were reconstructed by maximum parsimony, and maximum likelihood methods using the PAUP 4.0 program package (Swofford 1991). In addition, the consistency index, retention index, homoplasy index, and recalculated retention index were calculated for the parsimony analysis using PAUP* version 4.0b10 (Swofford, 2001). The bacterial species *Escherichia coli* was used as an out group in parsimony analysis due to it being within the same class. For the parsimony analysis, heuristic searches were completed with addition sequence set at random, 100 replicates, and the branch-swapping algorithm set at tree-bisection-reconnect (TBR). A bootstrap analysis was completed using 100 replicates to assess support for the clades. For the maximum likelihood analysis, the General Time Reversible (GTR) plus Gamma model was determined as the best maximum likelihood algorithm by the FindModel website (http://www.hiv.lanl.gov/content/sequence/findmodel/findmodel.html). For the maximum likelihood 10 heuristic searches were completed with addition sequence set at random, with branch-swapping set at neighbor joining. A bootstrap analysis was completed using 100 replicates to assess support for the clades.

The statistics for preliminary parsimony analysis are shown in Table 2. The analysis resulted in 1 tree of length 470, Consistency Index (CI) excluding uninformative characters of 0.3253, and Retention Index (RI) of 0.4302.

Table 2: Parsimony and maximum likelihood phylogenetic analysis summary of novel isolate and related species using 16S rRNA gene sequence data.

Parameter	**Parsimony Analysis**
Total Characters	1543
Variable Characters	190
Informative Characters	96
Most Parsimonious Trees	1
Total Tree Length	382
CI excluding uninformative characters	0.3253
RI	0.6284
Homoplasy Index excluding uninformative characters	0.5749
Rescaled consistency index	0.3570
	Maximum Likelihood Analysis
Gamma distribution	0.39388
LNL value	-1350.370134

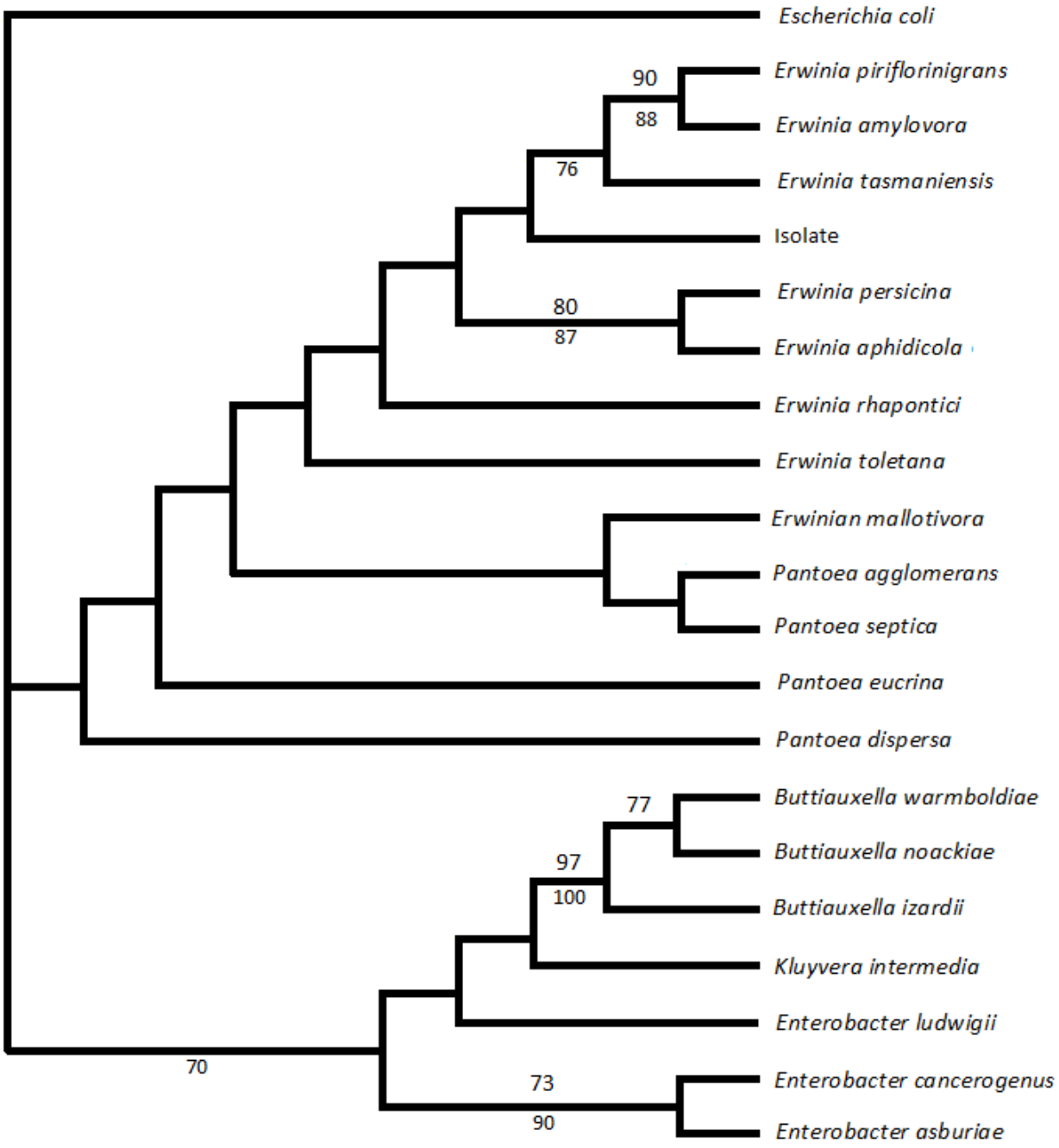

Figure 1: Initial parsimony and maximum likelihood tree identical topology for both methods based on 16S rRNA gene sequences showing the phylogenetic position of the isolate species and representatives of related taxa. Bootstrap values greater than 65% are included. Maximum likelihood values are above the line and parsimony values below lines. This tree represents the database records as of October 2016.

Section 2.1.2 Phylogenetic Analysis.

The twelve taxa that formed clades of less than 65% bootstrap support with the isolate in the preliminary tree constructions (Figure 1), were kept for further phylogenetic tree reconstruction and analysis (Table 3). The species *Erwinia persicina, Pantoea agglomerans,* and *Erwinia mallotivora* were then used to find additional closely related 16S rRNA gene sequences through SeqMatch under parameters of type strains, isolates, sequences greater 1200 base pairs, and high quality sequence. The taxa sampled, sequence similarity (as determined by RDP) and GenBank accession numbers are shown in Table 4.

The same methods for alignment and phylogenetic analysis were used. The branch stability of the parsimony phylogenetic tree was assessed by a thousand heuristic bootstrap replications. The branch stability of the maximum likelihood models was tested by a thousand bootstrap neighbor-joining replications.

Table 3: Twelve species that formed a clade with the isolate in preliminary tree re-constructions. Listed with species name, percent sequence similarity, and GenBank accession numbers.

Species Name	**Gen Bank Accession**	**Sequence Similarity**
Erwinia aphidicola	FN547376	92.7
Erwinia persicina	Z96086	92.5
Erwinia tasmaniensis	AM055716	91.5
Erwinia rhapontici	AJ233417	91.3
Erwinia toletana	FR870447	90.6
Pantoea dispersa	DQ504305	90.8
Erwinia amylovora	AJ233410	89.4
Pantoea eucrina	EU216736	89.4
Pantoea septic	EU216734	89
Erwinia piriflorinigrans	GQ405202	88.7
Erwinia mallotivora	AJ233414	88.1
Pantoea agglomerans	AJ233423	87.1

Table 4: List of additional species used for phylogenetic tree reconstruction and analyses. Species in bold denote the taxa used in preliminary tree reconstructions, that were then used to find additional related species based on 16S rRNA gene sequence similarity in the RDP SeqMatch program (as of October, 2016). Listed with each species name is the percent sequence similarity to either the isolate or a known species, and their GenBank accession numbers.

Species Name	Gen Bank Accession	Sequence Similarity	
		To Isolate	To Species
Erwinia persicina	**Z96086**	**92.5**	
Erwinia billingiae	JN175337		93.9
Pantoea agglomerans	**AJ233423**	**87.1**	
Pantoea vagans	EF688012		97.9
Pantoea brenneri	EU216735		96.7
Pantoea anthophila	EF688010		96.5
Pantoea eucalypti	EF688009		96
Pantoea conspicua	EU216737		95.8
Erwinia mallotivora	**AJ233414**	**88.1**	
Erwinia papaya	AY131237		93.6
Erwinia psidii	Z96085		92.1

Section 2.2.2: Phenotypic Analysis

Colony morphology and morphology and motility of the cells of the isolate were investigated using colonies grown on R2A media. Gram-staining was performed as described by Leboffe and Pierce (1997). Cell morphology was studied by negative staining using light microscopy (Leboffe and Pierce 1997), and by uranyl acetate stain transmission electron microscope imaging (K Grant, Wake Forest University). The latter was performed as follows: The isolate was incubated in R2B media for 24 hours. It was then suspended into water and adsorbed onto a carbon-coated formvar film which was attached to a metal specimen grid. Once the isolate was adsorbed onto the film surface, the excess sample was then blotted off to create a thin film of emulsion. The grid was then covered with a small drop of uranyl acetate stain solution. The emulsion was left on the grid for a few minutes and then blotted off. The sample was allowed to dry and was then examined in the transmission electron microscope at 13,000x magnification.

Growth characteristics, temperature, pH, and salinity, were determined using various R2A media plates. The temperature range for growth was determined by incubating cultures on R2A medium at 4, 12, 30, 37, and 50˚C. The optimum temperature for growth was determined by incubating on R2A plates at a range of 27℃ to 33℃. The isolate was also plated on R2A nutrient agar with pH levels of 3 through 9 to determine pH tolerance and on R2A nutrient agar with salinity levels of 1% to up 5% to determine salt preference and tolerance. A control culture on an R2A media plate with a pH 7 and 0% salinity was incubated at 25˚C. All plates were grown for 48 hours. The isolate was stabbed into medium with triphenyltetrazolium chloride to test for the presence of motility. This was allowed to grow for 72 hours.

Metabolic capabilities of the isolate were tested and all tests were performed in triplicate. Catalase activity was determined by exposing the culture to hydrogen peroxide and oxidase activity was determined using OxiStrips (Hardy Diagnostics). The isolate's ability to metabolize different substances was determined using BBL EnteroPluri Tubes (Becton Dickinson). Results were observed after 48 hours at 25˚C. Lipase activity was determined by plating the isolate on spirit blue agar that contained emulsified safflower oil. Inoculated plates were incubated at 30°C. The ability to hydrolyze gelatin was determined using a nutrient gelatin stab, and incubated at 25°C. The isolate's ability to ferment mannitol was tested using mannitol medium with phenol red incubated at 30°C. Results were observed after 48 hours. The ability of the isolate to convert tryptophan into indole was determined by using an indole with nitrate medium. The inoculated tube was incubated at 30°C for 48 hours. After 48 hours, Kovacs' Reagent was added to the tube and any color change was recorded. The isolate's ability to hydrolyze DNA was tested using DNAse media. Plates were incubated for 48 hours at 25°C, and were flooded with HCl and the absence or presence of halos was recorded. The isolate's caseinase activity was tested using casein medium. Plates were incubated for 48 hours at 25°C, and the absence or presences of halos was recorded. *Pseudomonas aeruginosa* was used as the control. The isolate's ability to reduce nitrate to nitrite was done by using a nitrate reduction test (Hardy Diagnostics). *Staphylococcus aureus* was used as the control for lipase, gelatinase, and DNAse activity. *E. coli* was used as the control for mannitol hydrolysis, indole production, and nitrate reduction. Recipes for lipase, DNase, gelatinase, phenol red with mannitol, caseinase, and gelatinase metabolic media were obtained from Leboffe & Pierce (2010).

The isolates susceptibility to the antibiotics: chloramphenicol (30 µg), furazolidone (300 µg), nalidixic acid (30 µg), oxytetracycline (30 µg), and tetracycline (30 µg) were tested by creating a lawn on R2A media, aseptically placing antibiotic disks on the media, and after 48 hrs of incubation, the isolates resistance or susceptibility was determined as described by Leboffe & Pierce (2010).

Section 2.2.3: Whole Genome Sequencing and Annotation

The isolate was sent to MR DNA, Inc. (Shallowater, TX) for whole genome sequencing and annotation. The library was prepared using a Nextera DNA Sample preparation kit (Illumina) following the manufacturer's user guide. The samples underwent the simultaneous fragmentation and addition of adapter sequences. These adapters were utilized during a limited-cycle (5 cycles) PCR in which unique indices were added to the sample. The libraries were then pooled in equimolar ratios of 2nM, and 10pM of the library pool was clustered using the cBot (Illumina) and sequenced paired end for 500 cycles using the HiSeq 2500 system (Illumina). In total twelve contigs were generated at sizes ranging from 7.8 KB to 2.8 MB. The genome was annotated by software within an Illumina workstation. Each of the twelve contigs was also analyzed by the program Basic Local Alignment Search Tool (BLAST) against the whole genome sequence of *E. billingiae* (https://www.ncbi.nlm.nih.gov/genome/1918) its closest relative as indicated by phylogenetic analysis. The BLAST program Multiple Sequence Alignment Viewer, was used to align the majority of generated contigs onto the *E. billingaie* reference genome (Genbank Assecion #JN175337).

The following proteins have been shown to be CSIs for the genus *Erwinia* (Adeolu et al. 2016): glutamate-cysteine ligase, DNA gyrase subunit B, LPS assembly protein LptD, thiol: disulfide interchange protein, two-component sensor histidine kinase, RNA helicase, tRNA pseudouridine (13) synthase TruD, glycine/betaine ABC transporter ATP-binding protein, transcriptional regulator, superoxide dismutase stationary phase inducible protein CsiE, and a hypothetical protein. These were all analyzed using BLAST for homologies in other species (Altschul et al. 1997).

Chapter 3: Results

Section 3.1: Isolation and Preliminary Classification of All Bacterial Cultures

The RDP Classifier program results are listed in Table 5. The program showed isolates 4, 7, 8, 11, 12, 13, 16, 18, and 21 had a 100% confidence of classification to the level of genus. Isolates 9, 14, 19, 20, and 27 each had a confidence of 92% or greater. Isolates 6, 10, 15, 17, 22, and 23 each had a confidence of 81% or below, with 6, 10, 13, 22, and 23 all classifying below 60% at the level of genus.

The most similar sequences as determined from the RDP SeqMatch program are listed in Table 6 and show the best relationships to the species level. Each of the isolates 7, 8, 12, 16, and 17 have a top sequence similarity ranging from 96.6 to 98.2% to organisms previously cultured or to a clone of 16S rRNA gene of related organisms. The isolates 4, 11, 18, 20, 21, 22, and 27 had a top sequence similarity ranging from 90.2 to 94.2%, while isolates 6, 10, 14, 15, and 19 all had top sequence similarities ranging from 83.4 to 85.6%, and isolates 9, 13, and 23 each had the highest similarity to database records ranging from 79.2 to 81.1%. Isolate 13 was chosen for further analysis.

Table 5: DNA sequence down to genus classification of bacterial isolates obtained from Oconaluftee Visitor Center and Kephart Prong, Great Smoky Mountains National Park, North Carolina, using the Ribosomal Database Project (RDP) Classifier software program, including isolation source and confidence interval value ("OR" is Oconaluftee River and "KP" is Kephart Prong).

Isolate	Isolation Source	Proposed Class	Proposed Order	Proposed Family	Proposed Genus
4	OR	Alphaproteobacteria*(100%)*	Sphingomonadales*(100%)*	Sphingomonadaceae*(100%)*	*Sphingomonas (100%)*
6	OR	Bacilli*(100%)*	Bacillales*(100%)*	Bacillaceae (66%)	*Falsibacillus* (38%)
7	OR	Actinobacteria*(100%)*	Actinomycetales*(100%)*	Micrococcineae*(100%)*	*Microbacterium (100%)*
8	OR	Alphaproteobacteria*(100%)*	Caulobacterales*(100%)*	Caulobacteraceae*(100%)*	*Asticcacaulis (100%)*
9	OR	Alphaproteobacteria*(100%)*	Sphingomonadales*(100%)*	Sphingomonadaceae[99%]	*Sphingomonas* (97%)
10	OR	Bacilli*(100%)*	Bacillales*(100%)*	Paenibacillaceae (95%)	*Paenibacillus* (44%)
11	OR	Flavobacteriia*(100%)*	Flavobacteriales*(100%)*	Flavobacteriaceae*(100%)*	*Flavobacterium (100%)*
12	OR	Actinobacteria*(100%)*	Actinomycetales*(100%)*	Sanguibacteraceae*(100%)*	*Sanguibacter (100%)*
13	OR	Gammaproteobacteria*(100%)*	Enterobacteriales*(100%)*	Enterobacteriaceae*(100%)*	*Erwinia (100%)*
14	OR	Actinobacteria*(100%)*	Actinomycetales*(100%)*	Micrococcaceae*(100%)*	*Micrococcus* (98%)
15	OR	Gammaproteobacteria *(100%)*	Pseudomonadales *(100%)*	Pseudomonadaceae (99%)	*Pseudomonas* (81%)
16	KP	Actinobacteria *(100%)*	Actinomycetales*(100%)*	Micrococcaceae*(100%)*	*Micrococcus (100%)*
17	KP	Betaproteobacteria *(100%)*	Burkholderiales*(100%)*	Oxalobacteraceae*(100%)*	*Rugamonas* (72%)
18	KP	Alphaproteobacteria *(100%)*	Caulobacterales*(100%)*	Caulobacteraceae*(100%)*	*Brevundimonas(100%)*
19	KP	Betaproteobacteria *(100%)*	Burkholderiales*(100%)*	Comamonadaceae*(100%)*	*Albidiferax* (97%)
20	OR	Actinobacteria *(100%)*	Actinobacteridae*(100%)*	Micrococcineae*(100%)*	*Plantibacter* (99%)
21	KP	Alphaproteobacteria *(100%)*	Rhizobiales*(100%)*	Methylobacteriaceae*(100%)*	*Methylobacterium*
22	OR	Betaproteobacteria *(100%)*	Burkholderiales*(100%)*	Comamonadaceae*(100%)*	*Albidiferax* (58%)
23	OR	Gammaproteobacteria *(100%)*	Pseudomonadales*(100%)*	Pseudomonadaceae (96%)	*Pseudomonas* (54%)
27	KP	Actinobacteria *(100%)*	Actinomycetales*(100%)*	Micrococcaceae*(100%)*	*Kocuria* (92%)

Table 6: DNA sequence matches of bacterial isolates obtained from Oconaluftee Visitor Center and Kephart Prong, Great Smoky Mountains National Park, North Carolina, using the Ribosomal Database Project (RDP) software program SeqMatch and including GenBank accession numbers for resulting matches with percentages indicating the extent of match of the GSMNP isolate to each isolate in the RDP database as of October 2015.

Isolate	SeqMatch Type Strain Matches (%)	GenBank Accession Number
4	*Sphingomonas oligophenolica* (91.9)	AB018439
6	*Bacillus circulans* (85.6)	AY724690
7	*Microbacterium (four records* (96.8)	AB234026, AB234027, AB234025, AB86028
8	*Asticcacaulis benevestitus* (97.7)	AM087199, AJ227758
9	*Sphingomonas formosensis* (79.2)	HM193517
10	*Paenibacillus (two records) (84.5)*	AB045094, EU558281, EU558284
11	*Flavobacterium aquidurense* (93.9)	AM177392
12	*Sanguibacter keddieii* (98.2)	X79450
13	*Erwinia aphidicola* (92.7)	FN54376
14	*Micrococcus (three records)* (83.4)	AJ536198, EU005372, FJ214355
15	*Pseudomonas psychrophila* (85.6)	AB041885
16	*Micrococcus (3 records) 97.6*	EU005372, FJ214355, AJ536198
17	*Rugamonas rubra* (96.6)	HM038005
18	*Brevundimonas aurantiaca* (92.1)	AJ227787
19	*Albidiferax ferrireducens* (85.4)	CP000267
20	*Plantibacter auratus* (93.8)	AB177868
21	*Methylobacterium cerastii* (90.4)	FR733885
22	*Albidiferax ferrireducens* (90.2)	CP000267
23	*Pseudomonas psychrophila* (81.1)	AB041885
27	*Kocuria palustris* (94.2)	Y16263

Section 3.2: Classification and Characterization of Bacterial Species

Section 3.2.1 Phylogenetic Analysis

The statistics for phylogenetic analyses are shown in Table 7. The parsimony analysis resulted in one tree of length 313, CI excluding uninformative characters of 0.5116, and RI of 0.7267. The ML, and parsimony strict consensus trees had the same topology and is shown in Figure 2, with bootstrap values above 65% given above (ML) and below (parsimony) the branches. The bootstrap analysis for both ML and parsimony have a third of the internal nodes on the tree showing high levels (BS values >90%) of support. These values in addition to the topological agreement of the ML and parsimony trees show strong support for the relationships shown on Figure 2. There is strong support that the isolate is closely related to *Erwinia billingiae* (BS 100%). The isolate and *E. billingiae,* were sister to a clade with the taxa *Erwinia piriflorinigrans, E. amylovora, and E. tasmaniensis* (BS<65%). The *Erwinia* taxa including *E. piriflorinigrans, E. amylovora, E. tasmaniensis, E. billingiae, E. aphidicola, E. persicina, E. rhapontici*, and *E. toletana* emerged as a clade with the isolate at 71%/69% BS (Clade A). Clade A occurred as sister clade (<65% BS) to the *Pantoea* clade (Clade B) made up of *P. eucalypti, P. anthophila, P. vagans, P. brenneri, P. conspicua,* and *P. agglomerans* (Clade B). *Pantoea* and *Erwinia* are not monophyletic in my analyses, suggesting the taxonomy of the group needs further study.

Table 7: Parsimony and maximum likelihood phylogenetic analysis summary of novel isolate and related species using 16S rRNA gene sequence data.

Parameter	**Parsimony Analysis**
Total Characters	1543
Variable Characters	160
Informative Characters	106
Most Parsimonious Trees	1
Total Tree Length	313
CI excluding uninformative characters	0.5116
RI	0.7267
Homoplasy Index excluding uninformative characters	0.4884
Rescaled consistency index	0.4342
	Maximum Likelihood Analysis
Gamma distribution	0.04
LNL value	-3060.233515

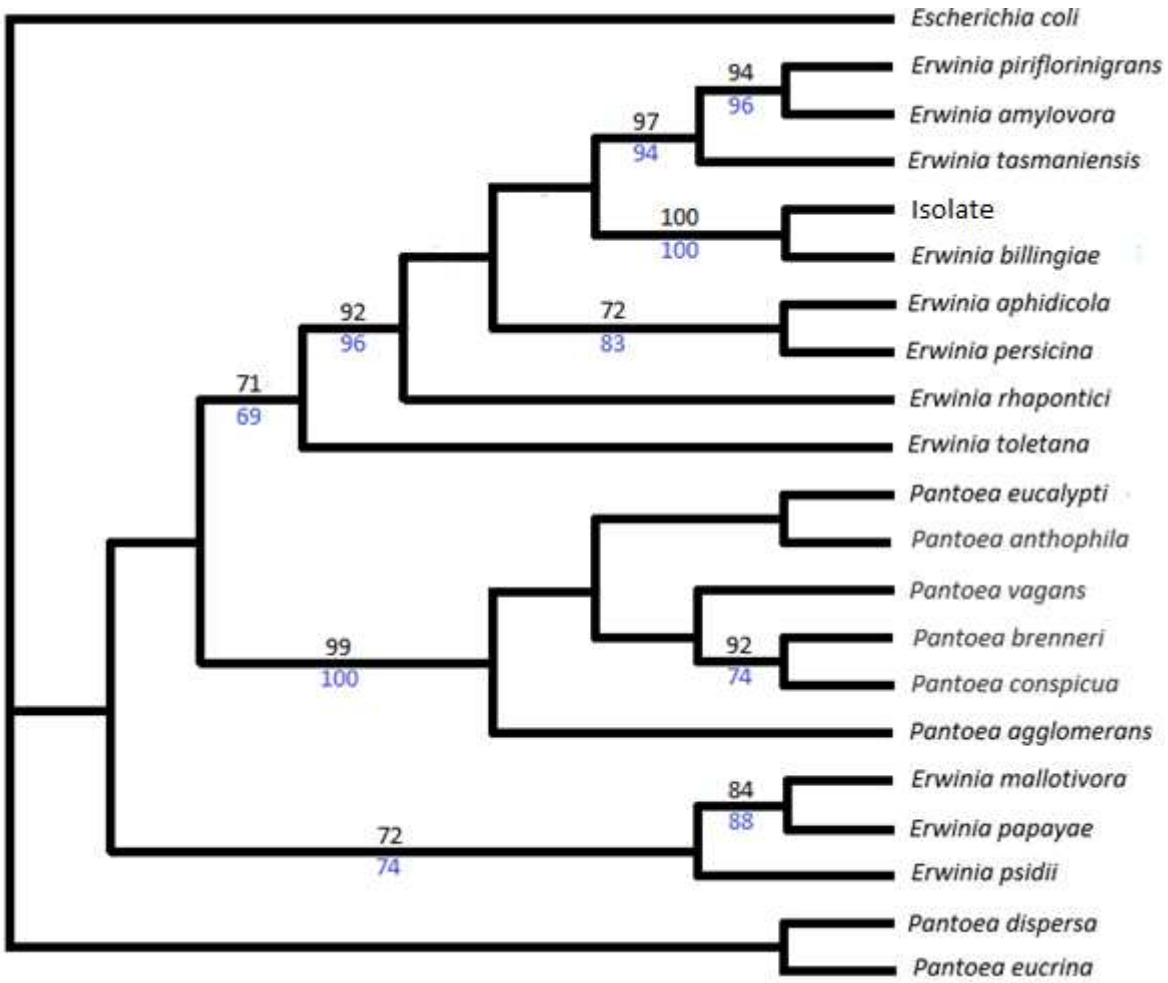

Figure 2: Parsimony and maximum likelihood using GTR with gamma tree based on 16S rRNA gene sequences showing the phylogenetic position of the novel isolate species and representatives of related taxa. Bootstrap values greater than 65% are included. Maximum likelihood values are above the line in black and parsimony values are below the lines in blue.

Section 3.2.2: Phenotypic Analysis

The isolate's cells were Gram-negative, short and rod shaped, 0.8–1.0 μm in length and 0.5 μm in diameter. Colonies were 1-2 mm in diameter, had a dull-translucent consistency, a circular shaped appearance, an undulate edging, a flat to convex elevation, and milky white coloring after 2 days of incubation on R2A media. Colonies would oftentimes form a mucoid

layer after 2 or more days of incubation. On R2A, the isolate could grow between 4 and 32°C with optimum growth between 27 and 30℃, at a pH range from 4.0 to 7.0, and it could tolerate up to 3% NaCl concentrations (Table 8). In the biochemical tests, the isolate was positive for the following: catalase activity; utilization of the sugars glucose, arabinose, lactose, and mannitol, and was capable of nitrate reduction (Table 9). These diagnostic characteristics of the isolate were then compared to results in the literature for *Erwinia amylovora, Erwinia billingiae, Erwinia tasmaniensis,* and *Erwinia piriflorinigrans,* chosen due to their forming a monophyletic clade with the isolate in phylogenetic tree reconstruction (Table 9). The isolate tested positive for motility and an electron micrograph image confirmed presence of flagella (Figure 3).

Table 8: Growth parameters for the bacterial isolate obtained from Kephart Prong, Great Smoky Mountains National Park, North Carolina, including ranges that permitted growth on R2A adjusted to various levels of NaCl, pH, and incubated at varying temperatures as well requirement for O_2. (A "+" means one or more tests yielded growth and a "-" means no tests yielded growth.

Temperature	
4°C	+
25°C	+
30°C	+
37°C	-
50°C	-
pH	
3	-
4	+
5	+
6	+
7	+
8	-
9	-
NaCl	
0%	+
1%	+
3%	+
4%	-
Oxygen	
20%	+
0%	-

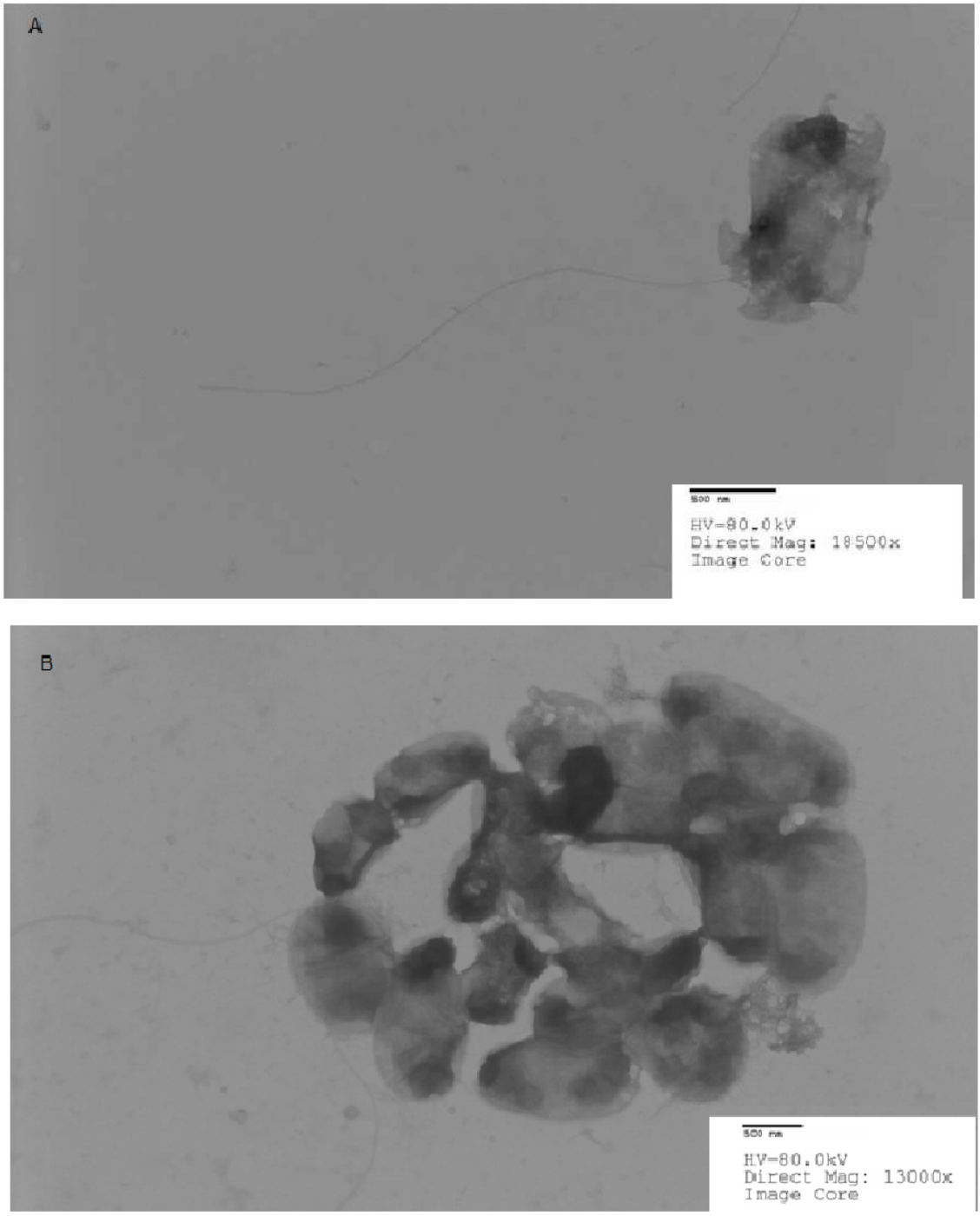

Figure 3: (A) Transmission electron micrograph image of the isolate and flagella and (B) transmission electron micrograph image, showing short rod shaped dimensions at 0.8–1.0 µm in length and 0.5 µm in diameter.

Table 9: Diagnostic characteristics of isolate and closely related species. A "+" means 90% or more of the strains are positive, a "-" means 10% or less of the strains are positive, a "d" means 11-89% of the strains are positive, and "nd" stands for not determined. Characteristics for related species gathered from Bergey's Manual (Hauben & Swings, 1998).

Characteristic	**Isolate**	***E. amylovora***	***E. billingiae***	***E. tasmaniensis***	***E. piriflorinigrans***
Growth in 5% NaCl	-	nd	nd	+	nd
Growth in 36℃	-	-	-	-	nd
$NO_3^- \rightarrow NO_2^-$	+	-	+	-	-
Production of indole	-	-	-	nd	+
Glucose fermentation	+	+	+	+	+
Lysine decarboxylation	-	nd	-	-	-
Ornithine decarboxylation	-	nd	nd	-	-
Sulfur reduction to H_2S	-	-	nd	-	nd
Adonitol fermentation	-	-	-	-	+
Lactose fermentation	d	nd	nd	nd	
Arabinose fermentation	+	d	nd	+	+
Sorbitol fermentation	-	d	nd	-	+
Production of acetoin	-	+	+	nd	+
Dulcitol fermentation	-	-	nd	-	-
Phenylalanine deaminase	-	-	nd	nd	nd
Urea hydrolysis	-	-	-	-	-
Citrate utilization	-	+	-	+	+
Oxidase	-	-	-	-	-
Catalase	+	+	+	+	+
Gelatinase	-	+	-	-	-
Lipid hydrolysis	d	nd	nd	nd	nd
Caseinase	-	-	nd	nd	nd
Mannitol hydrolysis	+	nd	+	nd	+
DNAse	-	nd	nd	nd	nd
Sensitivity to antibiotics					
Chloramphenicol (30 µg)	+	+	nd	nd	nd
Furazolidone (300 µg)	-	+	nd	nd	nd
Nalidixic acid (30 µg)	-	+	nd	nd	nd
Oxytetracycline (30 µg)	-	+	nd	nd	nd
Tetracycline (30 µg)	+	+	nd	nd	nd

Section 3.2.3: Whole Genome Sequencing and Annotation

The Multiple Sequence Alignment Viewer program allowed the placement of ten of the twelve contigs on the *E. billingiae* genome (Figure 4), with contigs 7 and 10 placed as two separate *E. billingiae* plasmids. Contig lengths, BLAST query coverage percentages, and coordinate placements of contigs onto to the *E. billingiae* genome are listed in Table 10. Sequence alignment analyses using BLAST show results for the twelve CSIs all have 100% matches with *E. billingiae* (Table 11).

A total of twelve contigs were able to be assembled from the sequencing data obtained for our isolate, with an average G-C content of 55.01%. The annotated genome came back with over 4,800 genes, with 30% being putative or hypothetical proteins. Out of the normally occurring 20 tRNAs, 18 were found in the genome, with a synthase protein for isoleucine present instead of that of the related tRNA. Proteins for chloramphenicol, and tetracycline resistance were found as expected based on our antibiotic sensitivity testing which showed isolate 13 is indeed resistant to both chloramphenicol and tetracycline. Additional proteins of interest are listed in Table 12.

Table 10: Contig lengths, BLAST query cover percentages, and coordinate placements of contigs to the *E. billingiae* genome. Data taken from BLAST and the multiple sequence alignment viewer program.

Contig Number	**Sequence Length (bp)**	**Percent Query Cover (%)**	**Sequence Start**	**Sequence End**
11	134,113	100	32,016	196,310
2	81,960	100	203,201	282,837
8	38,076	100	288,583	324,115
21	448,169	81	330,841	709,300
17	9,140	11	704,704	706,265
1	7,511	50	708,805	711,037
9	290,839	98	741,853	1,000,426
20	2,845,378	92	1,013,348	3,839,211
13	699,013	95	3,849,602	4,558,087
22	551,354	94	4,569,204	5,099,485

Table 11: Top BLAST results for the twelve conserved signature indels (CSIs) for genus *Erwinia.* Each protein matched 100% with the species *Erwinia billingiae.*

Matched Protein Description	**Max Score**	**Total Score**	**Query cover**	**E value**	**Ident**	**Accession**
Glutamate--cysteine ligase	1103	1103	100%	0.0	100%	WP_041692105.1
DNA gyrase subunit B	1662	1662	100%	0.0	100%	WP_013204563
LPS assembly protein LptD	1651	1647	100%	0.0	100%	WP_041691901
Thiol:disulfide interchange protein DsbA precursor	428	428	100%	8e-152	100%	WP_013200065
Two-component sensor histidine kinase	950	950	100%	0.0	100%	WP_013202973
RNA helicase	1291	1291	100%	0.0	100%	WP_013204010
Hypothetical protein	690	690	100%	0.0	100%	WP_013204009
tRNA pseudouridine(13) synthase TruD	706	706	100%	0.0	100%	WP_013203541
Glycine/betaine ABC transporter ATP-binding protein	805	805	100%	0.0	100%	WP_013203460
Transcriptional regulator	563	563	100%	0.0	100%	WP_013203422
Superoxide dismutase	425	425	100%	1e-150	100%	WP_013204530
Stationary phase inducible protein CsiE	857	857	100%	0.0	100%	WP_013203408

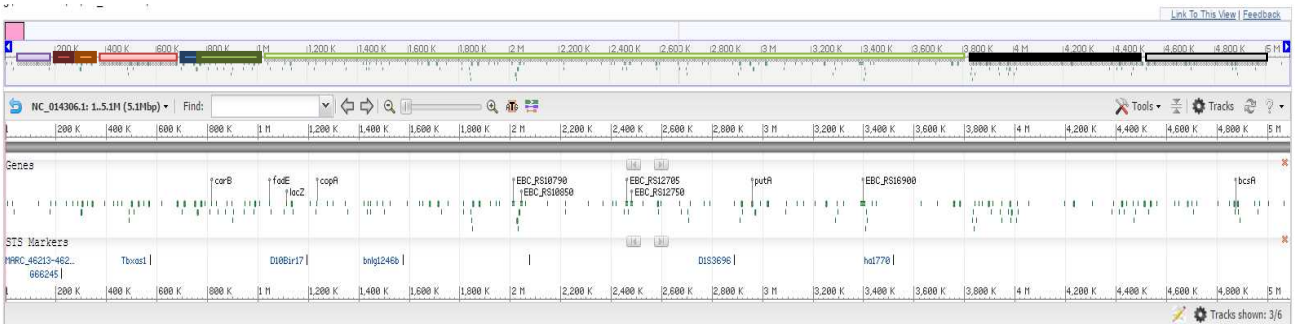

Figure 4: Placements of generated contigs of the isolate to the *Erwinia billingiae* complete genome. Each contig was aligned using BLAST and sequences were placed with the multiple sequence alignment viewer. Each contig is represented by a colored bar, placed below the top number line contig 11 (purple), contig 2 (red), contig 8 (orange), contig 21 (pink), contig 1 (blue), contig 9 (green), contig 20 (light green), contig 13 (black), and contig 22 (grey). Contig 17 was not placed because is likely is not a separate contig, but a small piece of contig 21.

Table 12: List of interesting or unusual groups of genes found within the annotated genome of the isolate.

Gene or protein types/functions	Total genes present
Drug resistance	65
Iron transport	25
Phage related	50
Flagella related proteins	39
Transposable element	8
Plasmid	17
Metal resistance	13
Pathogenesis related protein	1

Chapter 4: Discussion

Section 4.1: Isolation and Preliminary Classification of Bacterial Species

Due to low amounts of glaciation, relatively low anthropogenic impacts, and a wide range of habitats GSMNP, has long been considered one of the world's largest biodiversity hotspots (Nichols & Langdon, 2007). The initial inventory of bacterial diversity within GSMNP has shown little to no overlap between stream and soil samples and that different culturing methodologies produced different views of diversity even within the same sample (Nichols & Langdon, 2007). The 20 sequenced isolates recovered in this study show support for high diversity of bacteria in GSMNP. In total, the twenty sequenced isolates represent four different phyla of bacteria, that of the Bacteroidetes (1 species), Actinobacteria (6 species), Firmicutes (2 species), and Proteobacteria (11 species). Since sequence divergence values of ≥3% are considered to be strong evidence that the organisms are not related at the species level, after initial sequencing, fifteen of the twenty isolates may represent novel species within their respective genera (Stackebrandt & Goebel, 1994). Eight isolates (6, 9, 10, 13, 14, 15, 19, and 23) all have a top sequence similarity match ranging from 79.2 to 85.6%. These could potentially represent newly discovered genera, families, or orders, showing potential to further classify, characterize, and eventually be named as a novel species within a newly discovered genus or even a higher rank.

At a broader scope, the work of the ATBI in GSMNP over a fifteen-year period has cultured and/or sequenced thirteen different bacterial phyla, and potentially detected an estimated 800 distinct novel species within the park (O'Connell personal communication). The trends in the data show that the phylum Acidobacteria makes up the highest percentage of the collections

followed by the phylum Proteobacteria behind it. In the stream samples taken at both Oconaluftee Visitor Center and Kephart Prong, five different phyla were represented with the highest proportions being Proteobacteria followed closely by Actinobacteria and Bacteroidetes, reflecting a similar trend to the findings of this study (O'Connell et al. 2007).

Since its introduction in 1934, the Baas-Becking tenet, "everything is everywhere, but the environment selects" has been the main biogeographic hypothesis in regards to microbial distribution (O'Malley, 2008). New technologies for rapid and cost effective DNA sequencing and synthesis, has led to further understanding of the drivers that refute the Bass-Becking tenet and further clarify what biogeography and ecology models hold true at both the macro and micro level (Sul et al. 2013, Ryšánek et al. 2015, Larkin et al. 2017, Lindh et al. 2017). GMNSP has long been recognized as a global biodiversity hotspot and center of endemism for many groups of organisms including salamanders, lichens, and land snails (Getz and Uetz, 1994, Petranka, 2008, Lendermer and Allen 2017). With the high degree of endemism within GSMNP on the macro-organism scale, the same could be true in regards to microorganisms. Examining the microbial biodiversity in additional national parks, or in areas with high anthropogenic impacts may indicate if the high rates of microbial endemics within the park are due to GSMNP's unique ecology rather than the combination of the high quantity of microbial organisms/diversity and the relatively low number of microbial systematic studies.

Section 4.2 Classification and Characterization of Bacterial Species

Phenotypic, phylogenetic, and whole genome analyses indicate the studied isolate to be a strain of the species *Erwinia billingiae.* It is Gram negative, has rod-shaped cells, is catalase

positive and oxidase negative, and its inability to utilize adonitol, casein, and phenylalanine (Table 9, Figure 3) are all universal traits to the *Erwinia* genus (Hauben & Swings, 1998). The isolate's abilities to reduce nitrate, utilize arabinose, and utilize glucose are all traits consistent with the species *E. billingiae.* The topological agreement for both maximum likelihood and maximum parsimony phylogenetic reconstruction and the 100 percent branch support from each, show strong support for the isolate as a strain of *E. billingiae.* The median total length (Mb), protein coding gene count, and GC% of the *E. billingiae* genome is 5.48698, 4801, and 55.0709%, similar to that of isolate's genome of 5.278665, 4832, and 55.01% respectively (https://www.ncbi.nlm.nih.gov/genome/1918). The MSA Viewer program allowed the placement of the generated contigs onto either the *E. billingiae* genome or plasmid, and all twelve BLAST sequence alignments for the CSIs match 100% to the species.

The genus *Erwinia* was established in 1920 and was named after Erwin F. Smith. *Erwinia* was established from a group of gram-negative, non-sporulating, fermentative, peritrichous flagellated phytopathogenic bacteria (Hauben & Swings, 1998) and currently the genus is composed of 16 different species (Hauben & Swings, 1998). The relatively well studied species of the genus is *E. amylovora,* known for causing fire blight disease to such Rosaceous plants as pear, apple, raspberry, and several other ornamentals (Eastgate 2000). *E. billingiae* was reclassified from non-pigmented *E. herbicola* in 1998 (Mergaert, 1999), and has been isolated from Rosaceous plants in the United Kingdom as well as the Pacific Northwest. *E. billingiae* is often associated with plant pathogens, and are considered secondary invaders rather than primary pathogens. The presence of *E. billingiae* in GSMNP may have various macro-ecological implications.

The rosaceous family of plants is well represented within the park, with over seventy different species present (ATBI, 2017). It is possible that some precipitation or senescence event caused *E. billingiae* to be deposited from a leaf or other plant part and into the Oconaluftee River. Due to their lack of key virulence factors, and high efficiency in cell to cell communication, *E. billingiae* has been looked at as a biological control to fire blight; and has shown the ability to compete in growth and distribution of *E. amylovora* on flowers (Jakovljevic et al. 2008). At the present moment, fire blight is not one of the more actively monitored concerns of GSMNP management, and the presence of *E. billingiae* may be acting as a factor in preventing the disease. Iron transport proteins and siderophores specifically have been shown to function in two opposite ways during the infection process, both pathogenesis effectors in plant invasion and defense elicitors (Aznar et al. 2014). The twenty-five iron transport related proteins as well as the pathogenesis related protein found within the annotated genome of *E. billingiae* may help act as competitive inhibitors to *E. amylovora* and other pathogens.

As an epiphyte, *E. billingiae* may have potential for novel natural products in areas of carbon substrate utilization and its novel byproducts, antibiotic production, or biofilm formation (Smanski 2015, Larkin et al. 2017). Of the 4800 genes, some 30% are considered hypothetical proteins within the genome. Due to relatively low amounts of available resources on plant material, the environment may select for such epiphytic species as *E. billingiae* to be antibiotic producers able to defend their resource pool from competitors (Weiner 1996, Kinkel et al 2014). The 65 known drug resistant genes found within the classified isolate of *E. billingiae* genome shows the organism's potential abilities to ward off competition, and with 30% of its genome's function still unknown, novel antibiotics and natural products could be present. Further

characterization of the hypothetical proteins may lead to discovery of a novel natural product with industrial, medicinal, or ecological implications.

Genotypic data corroborates with phenotypic results, with many of the characteristics studied in Table 9 aligning with genes found in the annotated genome. For example, the isolate tested positive for both glucose and arabinose fermentation and multiple genes were found for both within the genome, as well as flagella present in the cell and the 39 flagella related proteins within the genome (Table 12). In some cases, the genome data runs in contrast to phenotypic tests, the three antibiotics furazolidone, nalidixic acid, and oxytetracycline the isolate showed zero sensitivity to, but had no genes present coding for their antibiotic resistance in the genome data. In contrast the isolate tested positive for both tetracycline and chloramphenicol sensitivity, while the genome data set showed two tetracycline resistance ribosome protection type genes, and one chloramphenicol acetyltransferase gene. These contrasting results show the importance of working with live cultures and analysis of their genomes.

Bacterial species evolve from three mechanisms: mutations, horizontal gene transfer, and transduction of phage DNA (Fortier et al. 2013, Bellanger et al 2014). The presence of plasmid and transposable element genes within the genome indicate *E. billingiae* has the ability to manipulate its genome in response to its environment through horizontal gene transfer. The fifty phage related genes indicate that at some point in its history, *E. billingiae* was able to uptake viral DNA into its genome. Further phylogenetic analyses using any of the plasmid, transposable element, or phage related genes may give further insight into *E. billingiae*'s molecular evolution.

At the time of original sequence analysis in October, 2016, the RDP program SeqMatch showed that the isolate was most closely related to that of *Erwinia aphidicola* at 92.7%

similarity, dissimilar enough to be classified as new species of bacteria. Sometime between October, 2016 and January, 2017 the RDP database was updated and found a 100% match between the isolate's 16S rRNA gene sequence and that of *Erwinia billingiae.* Though the original objective of this project was to isolate bacterial species from GSMNP and classify a new species, the discovery of a strain of *E. billingiae* has led to many interesting questions and insights. Many of the phenotypic tests done for the isolate, were the first results to be recorded for *E. billingiae* including antibiotic resistance. With up to millions of bacterial cells in river or soil samples what is the probability for finding this particular species again? Having been cultured from samples in the United Kingdom, the Pacific Northwest, and now GSMNP, what is the biogeographical history of *E. billingiae*?

Bacteria have become the cornerstone to many biological fields and the inspiration for many of these fields including molecular biology and taxonomy. The polymerase used for PCR was cultured from bacteria from thermal springs in Yellowstone National Park (Brock, Freeze 1969). Inexpensive gene and protein amplification occurs through microbial vectors, and many antibiotics are found through novel cultures (Zipperer et al. 2016). The gradual increase in knowledge of where microbial species occur, how they interact with one another, how they react to the environment itself, their physiological and metabolic capabilities, and their genomic flexibility has led to advancements in the fields of ecology, molecular evolution, biotechnology, biogeochemistry, and human health. Over its fifteen-year period, the ATBI has characterized and classified over 1700 prokaryotic species within GSMNP and of those, 31 or 1.75% are a part of the Enterobacteriaceae family with this study the first to find a species of *Erwinia* in the park. Understanding the likelihood of re-culturing and isolating *E. billingiae* or any other species

within the park would go a long way in elucidating what roles it plays within the environment in both a micro- and macro- ecological context.

Conclusion

A strain of the species *Erwinia billingiae* was isolated and classified through phenotypic, genotypic, and phylogenetic analysis from the Oconaluftee River, GSMNP. Through the classification process metabolic and antibiotic sensitivity characteristic tests were performed, many of which are the first time reported for *E. billingiae.* Out of the 1700 prokaryotic species so far recorded in the ATBI database, this was the is first instance of a genus of *Erwinia*, and one of 31 instances of a species from the Enterobacteriaceae family being found within the park.

Works Cited

Adeolu M, Alnajar S, Naushad S, and Gupta RS. 2016. Genome based phylogeny and taxonomy of the "Enterobacteriales": proposal for Enterobacterales ord. nov. divided into the families Enterobacteriaceae, Erwiniaceae fam. nov., Pectobacteriaceae fam. nov., Yersiniaceae fam. nov., Hafniaceae fam. nov., Morganellaceae fam. nov., and Budviciaceae fam. nov. International Journal of Systematic and Evolutionary Microbiology 66:5575-5599.

Aznar A, Chen N, Rigault M, Riache N, Joseph D, Desmaële D, Mouille G, Boutet S, Soubigou-Taconnat L, Renou J, Thomine S, Expert D, and Dellagi A. 2014. Scavenging Iron: A Novel Mechanism of Plant Immunity Activation by Microbial Siderophores, *Plant Physiology*, 164, 4, pp. 2167-2183.

Becraft E, Cohen FM, Kühl M, Jensen S, and Ward D. 2011. Fine-scale distribution patterns of *Synechococcus* ecological diversity in microbial mats of Mushroom Spring, Yellowstone National Park. Applied and Environmental Microbiology 77:7689-7697.

Bellanger X, Payot S, Leblond-Bourget N, and Guédon G. 2014, Conjugative and mobilizable genomic islands in bacteria: evolution and diversity, FEMS Microbiology Reviews, 38, 4, pp. 720-760.

Brock, TD, and Freeze H. 1969. *Thermus aquaticus* gen. n. and sp. n., a nonsporulating extreme thermophile. Journal of Bacteriology 98:289-297.

Cole JR, Wang Q, Fish JA, Chai B, McGarrell DM, Sun Y, Brown CT, Porras-Alfaro A, Kuske CR, and Tiedje JM. 2014. Ribosomal Database Project: Data and tools for high throughput rRNA analysis. Nucleic Acids Research. 42(Database issue):633-D642.

Cohan, FM. 2001. Bacterial species and speciation. Systematic Biology 50:513-524.

Curtis, TP, Sloan WT, and Scannell JW. 2002. Estimating prokaryotic diversity and its limits. Proceedings of the National Academy of Sciences 99:10494-10499.

Eastgate JA. 2000. *Erwinia amylovora*: The molecular basis of fireblight disease. Molecular Plant Pathology 1:325-329.

Ferris M, Kühl M, Wieland A, and Ward D. 2003. Cyanobacterial ecotypes in different optical microenvironments of a 68ºC hot spring mat community revealed by 16S-23S rRNA internal transcribed spacer region variation. Applied and Environmental Microbiology 69:2893-2898.

Fortier L, and Sekulovic O. 2013. Importance of prophages to evolution and virulence of bacterial pathogens. *Virulence*, 4, 5, pp. 354-365.

Fox GE, Magrum LJ, Balch WE, Wolfe RS, and Woese CR. 1977. Classification of methanogenic bacteria by 16S ribosomal RNA characterization. Proceedings of the National Academy of Sciences 74:4537–4541.

Garrido L, Sánchez O, Ferrera I, Tomàs N, and Mas J. 2014. Dynamics of microbial diversity profiles in waters of different qualities: Approximation to an ecological quality indicator. Science of the Total Environment 468-469:1154-1161.

Getz LL and Uetz GW. 1994. Species diversity of terrestrial snails in the Southern Appalachian Mountains, U.S.A. Malacological Review 27:61-74.

Groth I, Schumann P, Weiss N, Martin K, and Rainey FA. 1996. *Agrococcus jenensis* gen. nov., sp. nov., a new genus of actinomycetes with diaminobutyric acid in the cell wall. International Journal of Systematic Bacteriology 46:234–239.

Gupta RS. 2014. Identification of conserved indels that are useful for classification and evolutionary studies. Methods in Microbiology 41:153-182.

Hauben L and Swings J. 1998. Family 1 Enterobacteriaceae: Genus XIII *Erwinia.* Bergey's Manual of Systematic Bacteriology. 670-679. Williams &Wilkins, Baltimore Md.

Kim M, Hyun-Siok O, Sang-Cheol P, and Jongsik Chun. 2014. Towards a taxonomic coherence between average nucleotide identity and 16S rRNA gene sequence similarity for species demarcation of prokaryotes. International Journal of Systematic and Evolutionary Microbiology. 64:346-351

Kinkel LL, Schlatter DC, Xiao K, and Baines AD. 2014. Sympatric inhibition and niche differentiation suggest alternative coevolutionary trajectories among Streptomycetes. ISME Journal 8:249–256.

Koeppel A, Wertheim J, Barone L, Gentile N, Krizanc D, and Cohan F. 2013. Speedy speciation in a bacterial microcosm: new species can arise as frequently as adaptations within a species. ISME Journal 7:1080-1091.

Larkin A and Martiny A. 2017. Microdiversity shapes the traits, niche space, and biogeography of microbial taxa. Environmental Microbiology Reports 9:55-70.

Leboffe MJ and Pierce BE. 2010 Microbiology: Laboratory Theory & Application. Morton Publishing Company. Englewood, Colorado.

Lendermer J and Allen J. 2017. Reassessment of *Hypotrachyna virginica,* an endangered, endemic Appalachian macrolichen, and the morphologically similar species with which it has been confused. Proceedings of the Academy of Natural Sciences of Philadelphia 164:279-289.

Ling LL, Schneider T, Peoples AJ, Spoering AL, Engels I, Conlon BP, and Lewis K. 2015. A new antibiotic kills pathogens without detectable resistance. Nature, 517:455-459.

Lindh MV, Johanna S, Börje E, Casini M, Lundin D, Hugerth LW, Pinhassi J. 2017. Metapopulation theory identifies biogeographical patterns among core and satellite marine bacteria scaling from tens to thousands of kilometers. Environmental Microbiology 19:1222-1236.

Mahnert A, Moissl-Eichinger C, Berg G, Bik EM, and Mills DA. 2015. Microbiome interplay: Plants alter microbial abundance and diversity within the built environment. Frontiers in Microbiology 6:1-11.

McFall-Ngai MM. 2015. Giving microbes their due-animal life in a microbially dominant world. Journal of Experimental Biology 218:1968-1973.

Mesbah M, Premachandran U, and Whitman WB. 1989. Precise measurement of the G+C content of deoxyribonucleic acid by high-performance liquid chromatography. International Journal of Systematic Bacteriology 39:159–167.

Newton RJ, Jones SE, Helmus MR, and McMahon KD. 2007. Phylogenetic ecology of the freshwater Actinobacteria acI lineage. Applied and Environmental Microbiology 73:7169-7176.

Nichols, BJ, and Langdon KR. 2007. The Smokies All Taxa Biodiversity Inventory: History and Progress. Southeastern Naturalist 6:27-34.

Ndaw SM, Gama-Rodrigues AC, Gama-Rodrigues EF, Sales KN, and Rosado AS. 2009. Relationships between bacterial diversity, microbial biomass, and litter quality in soils under different plant covers in northern Rio de Janeiro State, Brazil. Canadian Journal of Microbiology 55:1089-1095.

O'Connell SP, York EA, Collins MB, Rosbach DT, Black KR, and Haney WB. 2007. An Initial Inventory of Bacteria Found within the Soils and Waters of Great Smoky Mountains National Park. Southeastern Naturalist, *6*(1), 57-72.

O'Malley MA. 2008. ‘Everything is everywhere: But the environment selects': Ubiquitous distribution and ecological determinism in microbial biogeography. Studies in History and Philosophy of Biological and Biomedical Sciences 39:314-325.

Petranka, JW. 1998. Salamanders of the United States and Canada. Smithsonian Institution Press, Washington D.C.

Quince C, Curtis TP, Sloan WT. 2008. The rational exploration of microbial diversity. ISME Journal 2:997-1006.

Ramos E, Ramírez-Bahena M, Valverde A, Velázquez E, Zúñiga D, Velezmoro C, and Peix A. 2013. International Journal Systematic and Evolutionary Microbiology 63:1834-1839.

Rappé MS, and Giovannoni SJ. 2003. The uncultured microbial majority. Annual Review Of Microbiology 57:369-394.

Ryšánek D, Kristýna H, and Pavel Š. 2015. Global ubiquity and local endemism of free-living terrestrial protists: phylogeographic assessment of the streptophyte alga Klebsormidium. Environmental Microbiology 17:689-698.

Seldin L and Dubnau D. 1985. Deoxyribonucleic acid homology among *Bacillus polymyxa*, *Bacillus macerans*, *Bacillus azotofixans*, and other nitrogen-fixing *Bacillus* strains. International Journal of Systematic Bacteriology 35:151–154.

Sloan WT, Woodcock S, Lunn M, Head IM, and Curtis TP. 2007. Modeling taxa-abundance distributions in microbial communities using environmental sequence data. Microbial Ecology 53:443–455.

Smith RJ, Paterson JS, Sibley CA., Hutson JL, and Mitchell JG. 2015. Putative effect of aquifer recharge on the abundance and taxonomic composition of endemic microbial communities. Plos One 10:1-17.

Stackebrandt E and Goebel BM. 1994. Taxonomic note: a place for DNA–DNA reassociation and 16S rRNA sequence analysis in the present species definition in bacteriology. International Journal of Systematic Bacteriology 44:846–849.

Stackebrandt E, Frederiksen W, Garrity GM, Grimont PD, Kämpfer P, Maiden MJ, Nesme X, Rosselló-Mora R, Swings J, Trüper HG, Vauterin L, Ward AC, and Whitman WB. 2002. Report of the ad hoc committee for the re-evaluation of the species definition in bacteriology. International Journal of Systematic and Evolutionary Microbiology 52:1043-1047.

Sul JW, Oliver TA, Hugh W, Ducklow L, Amaral-Zettler A, and Sogin ML. 2013. Marine bacteria exhibit a bipolar distribution. Proceedings of The National Academy of Sciences 110:2342-2347.

Swofford DL. 1991. PAUP: Phylogenetic Analysis Using Parsimony, Version 3.1 Computer program distributed by the Illinois Natural History Survey, Champaign, Illinois.

Van Belkum A, Yabuuchi E, and Yoshimasa K. 2006. Bergey's Manual of Systematic Bacteriology (Volume 2, Parts A–C, 2nd Edition). FEMS Immunology & Medical Microbiology, 234-272, 476.

Vandamme P, Pot B, Gillis M, de Vos P, Kersters K, and Swings J. 1996. Polyphasic taxonomy, a consensus approach to bacterial systematics. Microbiological Reviews 60(2), 407-438.

Whitman WB, Coleman DC, and Wiebe WJ. 1998. Prokaryotes: The unseen majority. Proceedings of the National Academy of Sciences 95:6578-6583.

Wiener P. 1996. Experimental studies on the ecological role of antibiotic production in bacteria. Evolutionary Ecology 10:405–421.

Woese CR and Fox GE. 1977. Phylogenetic structure of the prokaryotic domain: The primary kingdoms. Proceedings of the National Academy of Sciences 74:5088–5090.

Yamada Y, Aida K, and Uemura T. 1969. Enzymatic studies on the oxidation of sugar and sugar alcohol. V. Ubiquinone of acetic acid bacteria and its relation to classification of genera *Gluconobacter* and *Acetobacter*, especially of the so-called intermediate strains. Journal of General Applications of Microbiology 15:181-196.

Yarza P, Yilmaz P, Pruesse E, Glöckner FO, Ludwig W, Schleifer K, and Rosselló-Móra R. 2014. Uniting the classification of cultured and uncultured bacteria and archaea using 16S rRNA gene sequences. Nature Reviews Microbiology 12:635-645.

Zelles, L. 1997. Phospholipid fatty acid profiles in selected members of soil microbial communities. Chemosphere 35:275-294.

Zipperer A, Konnerth MC, Laux C, Berscheid A, Janek D, Weidenmaier C, Burian M, Schilling NA, Slavetinsky C, Marschal M, Willmann M, Kalibacher H, Schittek B, Brotz-Oesterhelt H, Grond S, Peschel A, Krismer B. 2016. Human commensals producing a novel antibiotic impair pathogen colonization. Nature 535:511-516.

Appendices

Appendix A: Complete 16S rRNA gene sequence

>isolate

GTAGCACAGAGAGCTTGCTCTCGGGTGACGAGTGGCGGACGGGTGAGTAATGTCTG
GGAAACTGCCTGATGGAGGGGGATAACTACTGGAAACGGTAGCTAATACCGCATAA
CGTCTTCGGACCAAAGTGGGGGACCTTCGGGCCTCACACCATCGGATGTGCCCAGAT
GGGATTAGCTAGTAGGTGGGGTAATGGCTCACCTAGGCGACGATCCCTAGCTGGTCT
GAGAGGATGACCAGCCACACTGGAACTGAGACACGGTCCAGACTCCTACGGGAGGC
AGCAGTGGGGAATATTGCACAATGGGCGCAAGCCTGATGCAGCCATGCCGCGTGTA
TGAAGAAGGCCTTCGGGTTGTAAAGTACTTTCAGCGGGGAGGAAGGCGATAAGGTT
AATAACCTTGTCGATTGACGTTACCCGCAGAAGAAGCACCGGCTAACTCCGTGCCA
GCAGCCGCGGTAATACGGAGGGTGCAAGCGTTAATCGGAATTACTGGGCGTAAAGC
GCACGCAGGCGGTCTGTCAAGTCAGATGTGAAATCCCCGGGCTTAACCTGGGAACT
GCATTTGAAACTGGCAGGCTAGAGTCTTGTAGAGGGGGGTAGAATTCCAGGTGTAG
CGGTGAAATGCGTAGAGATCTGGAGGAATACCGGTGGCGAAGGCGGCCCCCTGGAC
AAAGACTGACGCTCAGGTGCGAAAGCGTGGGGAGCAAACAGGATTAGATACCCTGG
TAGTCCACGCCGTAAACGATGTCGACTTGGAGGTTGTGCCCTTGAGGCGTGGCTTCC
GGAGCTAACGCGTTAAGTCGACCGCCTGGGGAGTACGGCCGCAAGGTTAAAACTCA
AATGAATTGACGGGGGCCCGCACAAGCGGTGGAGCATGTGGTTTAATTCGATGCAA
CGCGAAGAACCTTACCTGGCCTTGACATCCACGGAATTCGGCAGAGATGCCTTAGTG
CCTTCGGGAACCGTGAGACAGGTGCTGCATGGCTGTCGTCAGCTCGTGTTGTGAAAT
GTTGGGTTAAGTCCCGCAACGAGCGCAACCCTTATCCTTTGTTGCCAGCGAGTAATG
TCGGGAACTCAAAGGAGACTGCCGGTGACAAACCGGAGGAAGGTGGGGATGACGT
CAAGTCATCATGGCCCTTACGGCCAGGGCTACACACGTGCTACAATGGCGCATACA
AAGAGAAGCGAACTCGCGAGAGCAAGCGGACCTCACAAAGTGCGTCGTAGTCCGG
ATCGGAGTCTGCAACTCGACTCCGTGAAGTCGGAATCGCTAGTAATCGTAGATCAGA
ATGCTACGGTGAATACGTTCCCGGGCCTTGTACACACCGCCCGTCACACCATGGGAG
TGGGTTGCAAAAGAAGTAGGTAGCTTA

Appendix B: Whole Genome Sequences.

..\..\Amino Acid Whole Genome Contigs in Fasta.pdf

Appendix C: Alignment of the isolate 16S rRNA gene sequence in nexus format with closely related taxa.

..\..\SEQUENCES\Alignment Nexus Format.pdf

Appendix D: Partial 16S rRNA gene sequences of bacterial isolates obtained from Oconaluftee Visitor Center and Kephart Prong, Great Smoky Mountains National Park, North Carolina.

..\..\SEQUENCES\Partial 16S rRNA gene sequences of bacterial isolates .pdf

Appendix E: MSA View and the BLAST score for each generated contig of the isolate's genome.

..\..\..\..\Downloads\MSA view and blast score of each contig.pdf